

ABSTRACT

Title of Document: A UNIFIED EVALUATION OF GLOBAL AND LOCAL FIT
PERFORMANCE UNDER DIFFERING TEST
CONSTRUCTION CONDITIONS AND MODEL
MISSPECIFICATIONS

Matthew Michael Gushta, Doctor of Philosophy, 2012

Directed By: Associate Professor André A. Rupp

Department of Human Development and Quantitative
Methodology [formerly Department of Measurement, Statistics,
and Evaluation]

Social scientists and researchers frequently use latent variable models to analyze the relationships between observed variables and latent variables representing the hypothesized constructs. The population, or true, model is not always known, resulting in a degree of misspecification in the relationships between variables in the model. Therefore, model- and item-fit statistics have been developed in order to provide evidence for the validity of a specific latent variable model.

Conditions for mathematical equivalence between two popular latent variable modeling methods, confirmatory factors analysis (CFA) and item response theory (IRT), have been established, availing the researcher and practitioner of a variety of model- and item-fit indices. This dissertation employs a simulation design to examine the behavior of three model-fit indices (χ^2/df , RMSEA, and GDDM) and three item-fit indices ($S\text{-}\chi^2$, Modification Index, Wald Test) under various conditions of model misspecification and test design conditions. The results of this study show the empirically-derived cut points to out-perform the theoretical and suggested cut points when true models are estimated; these cut points are employed in subsequent analysis of misspecified models. In addition to examining the statistical power of each fit index to correctly reject the misspecified models, recommendations are made for the use of each fit statistic according to the model misspecification and test design conditions manipulated in the simulation study. Analysis of a real data set is provided as an illustration.

A UNIFIED EVALUATION OF GLOBAL AND LOCAL FIT PERFORMANCE
UNDER DIFFERING TEST CONSTRUCTION CONDITIONS AND MODEL
MISSPECIFICATIONS

By

Matthew Michael Gushta

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:

Associate Professor André A. Rupp, Chair
Associate Professor Robert G. Croninger
Professor Gregory R. Hancock
Professor Robert W. Lissitz
Professor Emeritus Robert J. Mislevy

© Copyright by
Matthew Michael Gushta
2012

Dedication

This dissertation would not have been possible if not for the boundless love and support provided by my wife, Sheri. While the writing is my own, this was a shared journey.

To my father who taught me to always be curious and that the story is ultimately more important than the facts.

Acknowledgements

Throughout this process, Dr. André A. Rupp has been an inspiring mentor and role model. With his kind yet insistent guidance, he helped me to refine my thinking and language, allowing me to grow and prosper as an independent researcher. His influence has positively affected not only this dissertation and my academic experience but also encouraged me to be ever attentive and precise as a professional. And, sometimes, to see the joy in the work.

I am thankful for the advice and instructional support provided by the faculty who guided me through the fundamentals and challenged me to rise above my own expectations. I am especially grateful to Dr. Robert J. Mislevy who, from the very first day, has been available, supportive, and patient throughout my academic career. Finally, I acknowledge Dr. Futoshi Yumoto for his collegial support, friendly criticisms, and coffee brewing mastery.

Table of Contents

Chapter 1 Introduction	1
1.1. Background	1
1.2. Model Specification	2
1.3. Estimation Frameworks.....	3
1.4. Organization	5
Chapter 2 Literature Review	7
2.1. Model Equivalence and Parameter Relationships	7
2.2. Q-Matrices.....	11
2.2.1 Structure and Function of Q-matrices	11
2.2.2 Illustrative Examples of Q-matrix Use in Previous Research	13
2.3. Summary of Notational Conventions	18
2.4. Properties of Model-Fit Indices.....	19
2.4.1 The χ^2/df Model-Fit Index	23
2.4.2 The RMSEA Index	24
2.4.3 The Generalized Dimensionality Discrepancy Measure.....	26
2.5. Properties of Item Fit Indices	28
2.5.1 The S- χ^2 Statistic.....	28
2.5.2 The Modification Index	29
2.5.3 The Wald Test.....	30
2.6. Summary	31
Chapter 3 Methods	33
3.1. Objective	33
3.1.1 Simulation Conditions	34
3.1.2 Data Generation.	47
3.2. Estimation Conditions	49
3.2.1 Model Misspecification	50
3.2.2 Fit Indices.....	51
3.2.3 Performance of Fit Statistics.....	53
3.3. Real Data Application	56
Chapter 4 Results of True Model Estimation	58
4.1. Estimation Issues.....	59
4.1.1 Results for MDIFF.....	61
4.1.2 Results for MDISC	61

4.1.3	Results for Inter-Factor Correlations	62
4.1.4	Results for Person Parameter Estimation.....	62
4.1.5	Summary of Estimation Issues.....	63
4.2.	Distributional Characteristics of Model Fit Indices	64
4.2.1	Results for χ^2/df	65
4.2.2	Results for RMSEA	69
4.2.3	Results for GDDM.....	73
4.3.	Distributional Characteristics of Item Fit Indices	76
4.3.1	Results for $S\text{-}\chi^2$	76
4.3.2	Results for Modification Indices.....	80
4.3.3	Results for Wald Test Statistics	84
4.4.	Estimation Bias for Type-I Error Rates under Theoretical Sampling Distributions.....	88
4.5.	Summary	97
Chapter 5 Results of Misspecified Model Estimation		100
5.1.	Estimation Issues	100
5.2.	Analysis of Model-Fit Indices under Model Misspecification	107
5.2.1	Distributional Characteristics of Model Fit Indices.....	107
5.2.3	Summary for Model-Fit Indices.....	117
5.3.	Analysis of Item-Fit Indices.....	119
5.3.3	Summary for Item-Fit Indices.....	138
5.4.	Synthesis of Model- and Item-Fit Performance Under Model Misspecification.....	140
5.4.2	Misspecification Correctly Detected by Model Fit Indices	141
5.4.3	Misspecification Not Detected by Model Fit Indices	146
5.5.	Summary	150
Chapter 6 Real Data Analysis		152
6.1.	Introduction	152
6.2.	Methods.....	153
6.3.	Results	156
6.3.1	Original Models	156
6.3.2	Revised Models.....	167
6.4.	Summary	174
Chapter 7 Discussion		178
7.1.	Summary of Key Findings	179

7.2. Considerations for Future Research	186
7.3. Conclusion.....	192
Appendix A Q Matrices	194
Appendix B Key Descriptive Statistics Under True Model Estimation	210
Appendix C Investigation into Convergence and Replication Issues	222
C.1. Non-Convergent and Heywood Cases	223
C.2. Determining the Optimal Number of Replications	223
C.3. Replications and the Two Factor Model	225
C.4. Replications and the Three-Factor Model.....	235
Appendix D Key Descriptive Statistics Under Misspecified Model Estimation	244
References	256

List of Tables

Table 2.1: <i>Between-Item Multidimensional Q-matrices for Wu and Adams (2006)</i>	15
Table 2.2: <i>Within-Item Multidimensional Q-matrix for Wu and Adams (2006)</i>	15
Table 2.3: <i>Q-matrices for Hartig and Höhler (2008)</i>	17
Table 3.1: <i>Simulation Design Summary</i>	35
Table 3.2: <i>MIRT Surface Plots for Each Item Type</i>	43
Table 3.3: <i>Summary of MIRT Item Parameters</i>	46
Table 4.1 <i>Convergence</i>	60
Table 4.2 <i>Descriptive Statistics for Root Mean-Squared Error and Average Bias of Key Parameters</i>	63
Table 4.3 <i>Selected Percentages of Variance for Empirically-Derived Model-Fit Cut Points Under True Model Specification</i>	65
Table 4.4 <i>Selected Percentages of Variance for Item-Fit Statistics by Simulation Condition Under True Model Specification</i>	76
Table 5.1 <i>Top 5 Percentages of Additional Replications Required when Estimated Models are Misspecified</i>	102
Table 5.2 <i>Descriptive Statistics for RMSE and Average Bias for Moderately Misspecified Models</i>	105
Table 5.3 <i>Descriptive Statistics for RMSE and Average Bias for Severely Misspecified Models</i>	106
Table 5.4 <i>Selected Percentages of Variance for Model-Fit Indices Under Model Misspecification, by Simulation Conditions</i>	108
Table 5.5 <i>Selected Percentages of Variance for Power of Model-Fit Statistics</i>	114
Table 5.6 <i>Descriptive Statistics for MDISC Values when Items were Correctly Specified or Alternate-Factored</i>	121
Table 5.7 <i>Types of Item Misspecification Present by Model Misspecification and Item Multidimensionality</i>	122
Table 5.8 <i>Selected Percentages of Variance for Item-Fit Statistics by Simulation Condition Under Model Misspecification</i>	123
Table 5.9 <i>Selected Percentages of Variance for Power of Item-Fit Statistics</i>	130
Table 6.1 <i>Q-matrices Resulting from 2- and 6-Dimensional Exploratory Factor Analysis and Cognitive Complexity</i>	155
Table 6.2 <i>Item Statistics Estimated for the 2-Dimensional Exploratory Factor Analysis Model</i>	157
Table 6.3 <i>Item Statistics Estimated for the 6-Dimensional Exploratory Factor Analysis Model</i>	159
Table 6.4 <i>Item Statistics Estimated for the Cognitive Complexity Model</i>	161

Table 6.5 <i>Design Appropriate Cut Points for the Grade 6 Mathematics Achievement Real-Data Analysis</i>	163
Table 6.6 <i>Model-Fit Estimates for the Original and Revised Models</i>	168
Table 6.7 <i>Item-Fit Values for the Revised EFA2 Model</i>	169
Table 6.8 <i>Item-Fit Values for the Revised EFA6 Model</i>	171
Table 6.9 <i>Item-Fit Values for the Revised COG Model</i>	173
Table 7.1: <i>Summary of Model and Item Fit Statistic Behaviour by Model Characteristics and Test Design Specifications</i>	181
Table A.1 <i>Q-Matrix for 2 Latent Factors and 12 Items according to Within-Item Multidimensionality</i>	194
Table A.2 <i>Q-Matrix for 2 Latent Factors and 12 Items according to Between-Item Multidimensionality</i>	195
Table A.3 <i>Q-Matrix for 2 Latent Factors and 24 Items according to Within-Item Multidimensionality</i>	196
Table A.4 <i>Q-Matrix for 2 Latent Factors and 24 Items according to Between-Item Multidimensionality</i>	197
Table A.5 <i>Q-Matrix for 2 Latent Factors and 36 Items according to Within-Item Multidimensionality</i>	198
Table A.6 <i>Q-Matrix for 2 Latent Factors and 36 Items according to Between-Item Multidimensionality</i>	200
Table A.7 <i>Q-Matrix for 3 Latent Factors and 12 Items according to Within-Item Multidimensionality</i>	202
Table A.8 <i>Q-Matrix for 3 Latent Factors and 12 Items according to Between-Item Multidimensionality</i>	203
Table A.9 <i>Q-Matrix for 3 Latent Factors and 24 Items according to Within-Item Multidimensionality</i>	204
Table A.10 <i>Q-Matrix for 3 Latent Factors and 24 Items according to Between-Item Multidimensionality</i>	205
Table A.11 <i>Q-Matrix for 3 Latent Factors and 36 Items according to Within-Item Multidimensionality</i>	206
Table A.12 <i>Q-Matrix for 3 Latent Factors and 36 Items according to Between-Item Multidimensionality</i>	208
Table B.1 <i>Key Descriptive Statistics for the χ^2/df Model-Fit Index Under True Model Estimation</i>	210
Table B.2 <i>Key Descriptive Statistics for the RMSEA Model-Fit Index Under True Model Estimation</i>	212
Table B.3 <i>Key Descriptive Statistics for the GDDM Model-Fit Index Under True Model Estimation</i>	213

Table B.4 <i>Key Descriptive Statistics for the $S\text{-}\chi^2/\text{df}$ Item-Fit Index Under True Model Estimation</i>	215
Table B.5 <i>Key Descriptive Statistics for Modification Index 1 Under True Model Estimation</i>	216
Table B.6 <i>Key Descriptive Statistics for Modification Index 2 Under True Model Estimation</i>	217
Table B.7 <i>Key Descriptive Statistics for Modification Index 3 Under True Model Estimation</i>	218
Table B.8 <i>Key Descriptive Statistics for Wald Test 1 Under True Model Estimation</i>	219
Table B.9 <i>Key Descriptive Statistics for Wald Test 2 Under True Model Estimation</i>	220
Table B.10 <i>Key Descriptive Statistics for Wald Test 3 Under True Model Estimation</i> ...	221
Table C.1 <i>2-Factor Model: Distributional and Key Indicators for Model Fit Indices Across Partition Sets</i>	226
Table C.2 <i>2-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets, Between-Item Multidimensionality</i>	230
Table C.3 <i>2-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets, Within-Item Multidimensionality</i>	233
Table C.4 <i>3-Factor Model: Distributional and Key Indicators for Model Fit Indices Across Partition Sets</i>	236
Table C.5 <i>3-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets</i>	240
Table D.1 <i>Key Descriptive Statistics for the χ^2/df Model-Fit Index Under Misspecified Model Estimation</i>	244
Table D.2 <i>Key Descriptive Statistics for the RMSEA Model-Fit Index Under Misspecified Model Estimation</i>	246
Table D.3 <i>Key Descriptive Statistics for the GDDM Model-Fit Index Under Misspecified Model Estimation</i>	248
Table D.4 <i>Key Descriptive Statistics for the $S\text{-}\chi^2/\text{df}$ Item-Fit Index Under Misspecified Model Estimation</i>	250
Table D.5 <i>Key Descriptive Statistics for Modification Index 1 Under Misspecified Model Estimation</i>	252
Table D.6 <i>Key Descriptive Statistics for Wald Test 1 Under Misspecified Model Estimation</i>	253

List of Figures

<i>Figure 3.1.</i> Kernel-smoothed density plots of the distributions of MDIFF values by Test Length and difficulty.....	45
<i>Figure 4.1.</i> Box-and-whisker plot for the χ^2/df ratio.	67
<i>Figure 4.2.</i> Cumulative distribution functions for the χ^2/df ratio.	68
<i>Figure 4.3.</i> Box-and-whisker plots for RMSEA.....	71
<i>Figure 4.4.</i> Empirical cumulative distribution functions for the RMSEA.	72
<i>Figure 4.5.</i> Box-and-whisker plots for GDDM.	74
<i>Figure 4.6.</i> Empirical cumulative distribution functions for the GDDM.	75
<i>Figure 4.7.</i> Box-and-whisker plots for $S\text{-}\chi^2$	78
<i>Figure 4.8.</i> Empirical cumulative distribution functions for the $S\text{-}\chi^2$	79
<i>Figure 4.9.</i> Box-and-whisker plots for the Modification Indices.	82
<i>Figure 4.10.</i> Empirical cumulative distribution functions for the Modification Index on latent factor 1.	83
<i>Figure 4.11.</i> Box-and-whisker plots for the Wald Tests.	86
<i>Figure 4.12.</i> Empirical cumulative distribution functions for the Wald Test on latent factor 1	87
<i>Figure 4.13.</i> Actual Type-I error rates for the χ^2/df ratio.	89
<i>Figure 4.14.</i> Actual Type-I error rates for the RMSEA.	91
<i>Figure 4.15.</i> Actual Type-I error rates for the $S\text{-}\chi^2$	94
<i>Figure 4.16.</i> Type I error rates for the Modification Index estimated against latent factor 1.....	95
<i>Figure 4.17.</i> Actual Type-I error rates for the Wald Test on latent factor 1.	96
<i>Figure 5.1.</i> Box-and-Whiskers Plots for χ^2/df under Model Misspecification.	110
<i>Figure 5.2.</i> Box-and-Whiskers Plots for RMSEA under Model Misspecification.....	111
<i>Figure 5.3.</i> Box-and-Whiskers Plots for GDDM under Model Misspecification.	112
<i>Figure 5.4.</i> Box-and-Whiskers Plots for Power of χ^2/df ratio.	115
<i>Figure 5.5.</i> Box-and-Whiskers Plots for Power of the RMSEA.....	116
<i>Figure 5.6.</i> Box-and-Whiskers Plots for Power of the GDDM.	117
<i>Figure 5.7.</i> Box-and-Whiskers Plots for MDISC Values when Items were Correctly Specified or Alternate-Factored.	121
<i>Figure 5.8.</i> Box-and-Whiskers Plots for the $S\text{-}\chi^2$ Under Model Misspecification.	125
<i>Figure 5.9.</i> Box-and-Whiskers Plots for Modification Index 1.....	126
<i>Figure 5.10.</i> Box-and-Whiskers Plots for Wald Test 1.	128
<i>Figure 5.11.</i> Box-and-Whiskers Plots for Power of the $S\text{-}\chi^2$	132

<i>Figure 5.12. Box-and-Whiskers Plots for Power of the Modification Index 1.</i>	134
<i>Figure 5.13. Box-and-Whiskers Plots for Power of the Modification Index 2.</i>	134
<i>Figure 5.14. Box-and-Whiskers Plots for Power of the Modification Index 3.</i>	135
<i>Figure 5.15. Box-and-Whiskers Plots for Power of the Wald Test 1.....</i>	136
<i>Figure 5.16. Box-and-Whiskers Plots for Power of the Wald Test 2.....</i>	137
<i>Figure 5.17. Box-and-Whiskers Plots for Power of the Wald Test 3.....</i>	137
<i>Figure 5.18. Power of item fit indices when χ^2/df ratio correctly indicates model misfit.</i>	143
<i>Figure 5.19. Power of item fit indices when RMSEA correctly indicates model misfit.</i>	144
<i>Figure 5.20. Power of item fit indices when GDDM correctly indicates model misfit. .</i>	145
<i>Figure 5.21. Power of item fit indices when χ^2/df ratio fails to indicate model misfit....</i>	147
<i>Figure 5.22. Power of item fit indices when RMSEA fails to indicate model misfit.....</i>	148
<i>Figure 5.23. Power of item fit indices when GDDM fails to indicate model misfit, for between-item multidimensional items.</i>	149
<i>Figure C.1. Distributional and key indicators for model fit indices, 2-factor models.....</i>	228
<i>Figure C.2. Distributional and key indicators for item fit indices, 2-factor models, between-item multidimensionality.....</i>	232
<i>Figure C.3. Distributional and key indicators for item fit indices, 2-factor models, within- item multidimensionality.</i>	234
<i>Figure C.4. Distributional and key indicators for model fit indices, 3-factor models.....</i>	238
<i>Figure C.5. Distributional and key indicators for item fit indices, 3-factor models.....</i>	242

Chapter 1

Introduction

1.1. Background

Latent variable models define a probabilistic relationship between observed responses to stimuli, such as test questions or items, and hypothesized constructs or abilities. Frequently employed among these are confirmatory factor analysis (CFA; see e.g., Brown, 2006; Gorsuch, 1983) and multidimensional item response theory (MIRT; see e.g., Ackerman, 1994; Embretson & Reise, 2000; Reckase, 2009) models which are both capable of representing specific relationships among observed responses and hypothesized latent constructs. These theoretical dependence relationships are represented using a priori constructed structures that serve to constrain the patterns of factor loadings or item discrimination parameters in CFA and MIRT models, respectively. That is to say that these structures constrain the associations between categorical response variables, or test items, and continuous latent variables, characterizing persons or examinees.

In the CFA literature, such dependence structures are referred to as the patterns of *factor loadings* in the measurement model and represented as the *factor loading matrix* Λ ; in psychometric research the structure may be referred to as a *Q-matrix*, which is often used to connect items to latent variables according to an a priori theory about task or item demands (Tatsuoka, 1983, 1984, 1990). The structures constructed to describe these connections are analogous to patterns of factor loadings specified in CFA models.

1.2. Model Specification

Specification of the model may correctly or incorrectly represent the underlying theory regarding the connections between observed and latent variables, regardless of whether the CFA or MIRT framework is employed. Correct model specification implies that the hypothesized model structure, as represented by the factor loading matrix or Q-matrix, matches that present in the population. Estimation of such a model may result in sample parameter estimates that differ from the population parameters – this does imply model misspecification but instead is the result of random sampling. Differences between sample estimates and population parameters reflect what Brown and Cudeck (1993) have termed “errors of estimation” and represent the degree of misfit between the sample and population model-implied covariance matrices.

Model misspecification can occur as a result of incorrect population distribution assumptions, use of an inappropriate link function in the item response function, missing data, unmodelled measurement error, failure to account for variable dependencies (Kaplan, 1990), or the misrepresentation of the theoretical association between observed and latent variables via the factor loading matrix or Q-matrix. Moreover, within a simulation context, correct model specification refers to the condition in which the estimating model matches the generating model; *misspecification of the measurement model* refers to models in which “(a) one or more parameters are estimated whose population values are zeros (i.e., an over-parameterized misspecified model), (b) one or more parameters are fixed to zeros whose population values are non-zeros (i.e., an under-parameterized misspecified model), or both” (p. 427, Hu & Bentler, 1998). Measurement model misspecification corresponds to misspecification of the pattern of factor loadings

in the estimating model – an incorrect Q-matrix. Subsequently, one important practical aspect of the successful and appropriate application of CFA or MIRT models includes the assessment of goodness-of-fit of the estimated models.

1.3. Estimation Frameworks

Generally, CFA is used to validate a hypothesized model structure or compare competing models. Under the CFA framework– and structural equation modeling (SEM) by extension – interpretations are typically made of the model as a whole; hence *global*- or *model-fit* statistics. From this foundation, statistics and methods have arisen to test the goodness-of-fit of estimated CFA models against absolute criterion, null models, and competing models. Additional statistics have been developed to detect and suggest modifications in the model that would improve goodness-of-fit. These goodness-of-fit indices have been shown to be differentially sensitive to types of misfit such as under-factoring, over-factoring, and misspecification of the measurement model (Fan & Sivo, 2005, 2007; Hu & Bentler, 1998).

Users of unidimensional item response theory (IRT) and MIRT, on the other hand, are typically concerned with the interpretation of specific observed variables or test items and the unobserved or latent ability of examinees. Stemming from the assumption that the IRT or MIRT model being applied is correct or valid, fit indices then describe the deviation of items or examinees from the given item response model. Therefore, few model-fit indices have been specifically developed for application under an IRT or MIRT framework; instead the focus has been on person- and item-fit indices. Item fit analysis describes model-data fit for each item by comparing model predictions to actual responses. The resulting statistics are useful in describing the functioning of the test in

terms of items and students, however, IRT models do not typically yield diagnostic information regarding model-fit such those provided when CFA models are estimated.

Fortunately, the equivalence between MIRT and CFA models has been established providing a number of assumptions are met (Kamata & Bauer, 2008; Takane & de Leeuw, 1987), which will be discussed further in Chapter 2. When these assumptions are met, IRT and CFA models yield parameters that are interchangeable after application of known transformation formulae (Takane & de Leeuw, 1987). Though these models differ in regards to the purpose for which they are typically been employed and the subsequent inferences made based on the results, statistical equivalence between models suggests that desirable features of both can be employed to explore and describe global, model-fit and local, item-fit.

Recently, research has been conducted regarding the application and behaviour of select model-fit indices adopted from the factor analytic framework within an IRT context (Harrell, 2009). However, there has been no research specifically investigating the implications of wide-spread measurement model misspecification on model-fit indices applied within a MIRT context. Further, research adjudging model-fit has been limited in scope, examining the effects of fixing or freeing only one or two loadings, and has failed to fully consider the effects of other aspects of the data that would be of interest in large-scale assessment, such as item difficulty (D. Jackson, personal communication, November 4, 2009).

The current study proposes a Monte Carlo simulation in the examination of model- and item-fit for data generated under conditions of equivalence between CFA and

MIRT models per Takane and de Leeuw (1987), to which various types of measurement model misspecification are applied under a range of varying test characteristics. These characteristics include varying sample size, varying item difficulty and item discrimination parameter specifications, varying dimensional correlations, and varying types of Q-matrices.

The results of this study will provide researchers with information about the performance of model- and item-fit indices under various item difficulty and discrimination conditions and the impact of potential measurement model misspecification. Specifically, it will inform researchers about which types of fit statistics designed and applied to equivalent CFA and MIRT models are most suitable to detecting different kinds of model misspecifications.

1.4.Organization

This dissertation follows a seven-chapter structure. Chapter 1 has introduced the concepts and background for the research. In Chapter 2, the conditions necessary for equivalence between CFA and MIRT models are described and literature describing and demonstrating the construction and use of Q-matrices are reviewed. Subsequent to a summary of notational conventions, an overview of the literature on the properties of model- and item-fit indices is provided. Chapter 3 describes the methodology applied in this dissertation after clearly stating the objectives in the form of research questions. In this chapter, the simulation and model estimation conditions are described and the methods of evaluation of the resulting estimates are detailed. Chapters 4 and 5 present the results of the simulation according to true model estimation or misspecified model estimation, respectively. Within each of these chapters, estimation issues and recovery of

person and item parameters are first examined, then the performance of the model- and item-fit statistics are separately described. In Chapter 4, theoretical and empirical cut points are described; in Chapter 5 power is demonstrated as resulting from the application of the empirical cut points. Chapter 5 concludes by synthesizing and summarizing the information provided by model- and item-fit results. With information about the performance of model- and item-fit statistics under various simulation conditions, the fit of various Q-matrices to a real data set is evaluated in Chapter 6. Lastly, Chapter 7 concludes with a summary of the key findings, theoretical and practical implications for these results, consideration for the limitations of the current research, and suggested topics for future research.

Chapter 2

Literature Review

This chapter first describes the conditions necessary to establish equivalence between the confirmatory factor analysis (CFA) and multidimensional item response theory (MIRT) frameworks, detailing specific assumptions and transformations required to be able to employ one or both frameworks in the study of measurement model misspecification. The definition of the Q-matrix and its role in the CFA and MIRT frameworks is described. To be applied under each of the frameworks in the evaluation of measurement model misspecification, the notions of model- and item-fit are described. This is followed by a review of the literature that is focused on the properties of those model-fit and item-fit indices that are identified as appropriate for detecting measurement model misspecification when estimating CFA and MIRT models.

2.1. Model Equivalence and Parameter Relationships

In educational and psychological measurement, two classes of models are commonly utilized for the purpose of relating multiple observed or manifest variables to one or more latent variables. The following sections describe the core ideas and results; more detailed descriptions can be found in sources such as Brown (2006) for CFA, Reckase (2009) for MIRT, as well as McDonald (1999) and Thissen and Wainer (2001), which describe the statistical and practical connections between these two modeling frameworks.

Factor analytic (FA; Gorsuch, 1983) models estimate patterns of covariation via linear relationships between the observed response variables and multiple latent variables when the observed variables are continuous. Item response theory models (IRT; e.g. Lord

& Novick, 1968; Embretson & Reise, 2000), on the other hand, define nonlinear relationships between the hypothesized latent variable and observed responses, which are assumed to be discrete. Specifically, multidimensional IRT models (MIRT; e.g., Ackerman, 1994) extend unidimensional IRT models to allow for more than one latent variable. Similarly, nonlinear factor analysis (NLFA) or item factor analysis (IFA) models (e.g., De Champlain, 1999; McDonald, 1999) attempt to overcome the technical issues presented when the data is non-continuous.

Fortunately, there has also been a large amount of research establishing the formal similarity between IRT and IFA approaches (Kamata & Bauer, 2001; Mislevy, 1986; McDonald, 1999; Takane & de Leeuw, 1987); the equivalence between one- and two-parameter IRT and CFA models has been established providing a number of assumptions are met (Kamata & Bauer, 2008; Takane & de Leeuw, 1987).

Specifically, under the two-parameter normal-ogive MIRT model (Reckase, 2009), the probability of a correct response to binary item j given abilities $\theta_1 \dots \theta_k$ is calculated as¹:

$$P(x_j = 1 | \mathbf{a}_j, b_j, \boldsymbol{\theta}) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-(z^2/2)} dz$$

where $z = \mathbf{a}_j(\boldsymbol{\theta}' - b_j)$, \mathbf{a}_j is a $1 \times k$ row vector of item discrimination parameters for item j , $\boldsymbol{\theta}$ is the row vector of k latent variable scores, and b_j is an item difficulty parameter. The general FA model, however, presumes continuous observed variables and is typically expressed as:

$$\mathbf{Y}^* = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

¹ Subscript i indexing students or examinees $1, \dots, N$ has been excluded from this section for clarity.

where \mathbf{Y}^* is the $j \times 1$ vector of observed continuous responses where j indexes items 1, ..., J ; $\boldsymbol{\tau}$ are the $j \times 1$ intercepts or threshold parameters; $\boldsymbol{\Lambda}$ is the $j \times k$ matrix of slopes or factor loadings where k indicates the number of latent factor scores and $k < j$; $\boldsymbol{\theta}$ is the $k \times 1$ vector of latent factor scores; and $\boldsymbol{\varepsilon}$ is the $j \times 1$ vector of random errors.

The general FA or CFA model assumes that errors are normally distributed as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a $j \times j$ diagonal matrix of variance in $\boldsymbol{\varepsilon}$, and that errors and latent factors are uncorrelated, $\text{cov}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) = 0$. The marginal distribution of the continuous observed response is assumed to follow

$$\mathbf{Y}^* \sim N(\boldsymbol{\tau}, \boldsymbol{\Sigma})$$

where the threshold parameters are usually assumed to be $\boldsymbol{\tau} = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ given the $k \times k$ inter-factor correlation matrix $\boldsymbol{\phi}$. The conditional distribution of \mathbf{Y}^* given $\boldsymbol{\theta}$ is

$$\mathbf{Y}^* | \boldsymbol{\theta} \sim N(\boldsymbol{\Lambda}\boldsymbol{\theta}, \boldsymbol{\Psi})$$

In order to use the common FA model to analyze dichotomous data similar to MIRT models, it is necessary to make the further assumption that y_j^* is an unobserved continuous response which is manifested by a dichotomized variable x_j given the following relationship:

$$x_j = \begin{cases} 1 & \text{if } y_j^* \geq \tau_j \\ 0 & \text{if } y_j^* < \tau_j \end{cases}.$$

Following the above assumptions, the marginal probability that $x_j = 1$ conditional on $\boldsymbol{\theta}$ is obtained under a CFA framework as

$$P(x_j = 1 | \boldsymbol{\theta}) = \int_{\tau_j}^{\infty} f(x_j = 1 | \boldsymbol{\theta}) dy_j = \Phi\left(\frac{\boldsymbol{\lambda}_j \boldsymbol{\theta} - \tau_j}{\psi_j}\right)$$

where ψ_j , residual or standard error for item j as the j^{th} diagonal from Ψ , can be alternately set to $\psi_j = 1.0$ or estimated as $\psi_j = \sqrt{1 - \lambda_j' \Phi \lambda_j}$ (Kamata & Bauer, 2008; McDonald, 1999).

Formal equivalence between CFA and MIRT models is, therefore, established given that the errors in ε are assumed to be normally distributed and independent, often referred to as the assumption of local independence, and that the latent factors in the CFA model are scaled to have a multivariate normal distribution, $\theta \sim \text{MVN}(0, \Sigma)$.

Slope and threshold parameters in the CFA framework are related to discrimination and difficulty in the IRT framework as² $\mathbf{a}_j = \frac{\lambda_j}{\psi_j}$ and $d_j = -\frac{\tau_j}{\psi_j}$. For

MIRT, item difficulty and discrimination values can be represented as scalars *MDIFF* and *MDISC* for each item (Reckase, 2009). When the means of the latent variables are set to $\theta_k = 0$ for scale identification, MDIFF represents the distance from the origin of the k -dimensional item response surface to the point of steepest slope; it is calculated as

$$\text{MDIFF} = -d_j / \sqrt{\sum_{k=1}^m a_{jk}^2}, \text{ and } d_j = -b_j \sqrt{\sum_{k=1}^m a_{jk}^2} \text{ when the alternate parameterization for}$$

difficulty has been used. Similar to the unidimensional b -parameter, positive MDIFF values indicate more difficult items while negative values indicate easier items.

The value of the multidimensional discrimination parameter, MDISC, is

$$\text{calculated as } \text{MDISC} = \sqrt{\sum_{k=1}^m a_{jk}^2} \text{ and represents the slope in the item response surface at}$$

the location indicated by MDIFF. Values of MDIFF and MDISC can be displayed

² Kamata and Bauer (2008) describe other parameterizations the use of reference indicators typical of FA research. The current paper focuses on the typical IRT practice of defining the latent trait as $\theta \sim N(0, 1)$.

graphically in item vector plots whereby the base of each i item vector is located at MDIFF and the angle, a_{j1} , with respect to coordinate axis 1, calculated as:

$$\cos a_{j1} = a_{j1} / \sqrt{\sum_{k=1}^m a_{jk}^2} . \text{ The length of each item vector is determined as MDISC.}$$

2.2. Q-Matrices

2.2.1 Structure and Function of Q-matrices

The original definition and description of Q-matrices was provided by Kikumi Tatsuoka (1983, 1984, 1990) as $k \times j$ incidence matrices where k indexes attributes and j indexes test items. In early applications, the Q-matrix was used to represent the specific operations that were necessary to successfully answer each item on a mathematics assessment; the specific operations included concepts or attributes like addition, subtraction, and multiplication. The Q-matrix was then utilized within a multidimensional classification framework to analyze student response data for the purpose of diagnosing "bugs" or difficulties with respect to one or multiple of the attributes.

Since that time, Q-matrices have been typically presented as $j \times k$ incidence matrices indicating specific requirements for test items, often corresponding to cognitive demands (see Rupp, Templin, & Henson, 2010 for state of the art applications and examples). Formally, element $q_{jk} = 1$ in the Q-matrix indicates that item j loads on / requires / measures attribute / latent factor / dimension k for a successful response, and $q_{jk} = 0$ indicates that item j does not load on / require / measure attribute / latent factor / dimension k . The unidimensional IRT model is a special case in which the Q-matrix is simply described as a column vector for which all entries are 1, indicating that the item

discrimination values are associated with the single latent factor and freely estimated. Further, the granularity of the attributes and resulting interpretations of the incidence elements of any Q-matrix is not limited to cognitive processes or other such fine differentiations but can be as broad or detailed as the substantive theory necessitates.

Under multidimensional CFA and MIRT models, items may demonstrate either *between-item multidimensionality* or *within-item multidimensionality* (Adams, Wilson, & Wang, 1997), respectively known as simple or complex structure in the CFA literature. Items demonstrating between-item multidimensionality conform to simple structure and are associated with a single latent factor; the Q-matrix row j contains only one element where $q_{jk} = 1$. Within-item multidimensionality, however, is used to describe items with complex loading structures; multiple entries of $q_{jk} = 1$ are present for item j .

From a statistical standpoint, the Q-matrix serves to clearly define the theorized associations between observed and latent variables. Item and person characteristics are subsequently reported with respect to the latent factors or attributes represented by the columns in the Q-matrix. Put differently, the Q-matrix serves to represent the constraints that are applied to certain model parameters for the purpose of representing substantive theory. Under the CFA and MIRT models, entries of the Q-matrix imply the pattern of fixed and freely estimated measurement model parameters. In this dissertation, the Q-matrix is used as a structural component in parametric latent variable models (i.e., CFA and MIRT models) where it serves to constrain the factor loadings (CFA) or item discrimination parameters (MIRT).

2.2.2 Illustrative Examples of Q-matrix Use in Previous Research

Before describing the simulation conditions and specific use of the Q-matrix in this dissertation, the following studies demonstrate instances where Q-matrices were applied, or could have been applied, to item response data. These examples highlight implicit or explicit application of Q-matrices within CFA and MIRT frameworks according to differing numbers of latent factors, or dimensions, and demonstrating between-item multidimensionality (simple structure) or within-item multidimensionality (complex structure).

The first study considered is that by Wu and Adams (2006) in which students' responses to mathematics problem solving tasks were explored. The authors first posed a four-dimensional problem solving framework based on three principles: (1) the latent factors or dimensions needed to be related to instructionally-relevant information and performance; (2) the dimensions must be associated with observable student behavior; and (3) test response data could be modeled and analyzed using available software. From these principles, four dimensions were defined as (1) reading/extracting all information from the question; (2) real-life and common sense approach to problem-solving; (3) mathematics concepts, "mathematisation", and reasoning; and (4) standard computational skills and carefulness in carrying out computations. Using these definitions, four different test forms were designed which comprised a total of 48 items, one-quarter of which were multiple-choice while the majority were polytomously scored. These test forms were administered to 951 grade 5 and 6 students in the suburbs of Sydney and Melbourne, Australia.

Item response data was modeled using the Random Coefficient Multinomial Logit Model (Adams, Wilson, & Wang, 1997) implemented in the ConQuest software (Wu, Adams, & Wilson, 1998) which estimates the partial credit model – a polytomous extension of the Rasch IRT model. Two different models were estimated which followed between-item multidimensionality (simple structure): the two-dimensional model grouped items as (1) heavy reading and (2+3+4) all others; the three-dimensional model grouped items as (1) heavy reading, (2) common-sense mathematics, and (3+4) all others. A unidimensional model was also estimated for the sake of comparison. Though Q-matrices were not provided for this study, general forms can be seen in Table 2.1.

Table 2.1:
Between-Item Multidimensional Q-matrices for Wu and Adams (2006)

Item Group	Uni-dimensional	2-Dimensional		3-Dimensional		
		Reading/ extracting	All Others	Reading/ extracting	Common Sense	All Others
Heavy Reading	1	1	0	1	0	0
...	1	1	0	1	0	0
Common Sense	1	0	1	0	1	0
...	1	0	1	0	1	0
Math Concepts	1	0	1	0	0	1
...	1	0	1	0	0	1
Computation	1	0	1	0	0	1
...	1	0	1	0	0	1

Tests of model deviance showed that the three-dimensional model fit best compared to the two-dimensional and unidimensional model. Additionally, a four-dimensional within-item multidimensional model was estimated according to the four specified dimensions plus additional factor loadings suggested by confirmatory factor analysis (these are not detailed in the paper); the general form of the four-dimensional Q-matrix is also presented in Table 2.2.

Table 2.2:
Within-Item Multidimensional Q-matrix for Wu and Adams (2006)

Item Group	Reading/ extracting	Common Sense	Math Concepts	Computation
Heavy Reading	1	*	*	*
...	1	*	*	*
Common Sense	*	1	*	*
...	*	1	*	*
Math Concepts	*	*	1	*
...	*	*	1	*
Computation	*	*	*	1
...	*	*	*	1

* Additional factor loadings not detailed by author.

When compared to the results of exploratory factor analysis (EFA), the authors found that the MIRT results produced interpretable student profile information while the EFA results were uninformative and prone to representing idiosyncratic disturbances in item features. While the inter-factor correlations for the multidimensional models suggest unidimensionality, ranging $\rho = [0.79, 0.95]$, the multidimensional model demonstrated better fit than the unidimensional model. Further, these values were shown to be comparable to those reported for the Programme for International Student Assessment (PISA; Adams & Wu, 2002).

In a second example, Hartig and Höhler (2008) modeled German, Austrian, and Italian students' responses to English reading and listening comprehension tests. Specifically, the authors were interested in whether between- or within-item multidimensional models resulted in different substantive implications, as demonstrated by the goodness-of-fit results and the patterns of factor loadings.

Two English as a foreign language tests were administered to 9557 grade 9 students: the reading comprehension test consisted of 46 multiple-choice items requiring students to decode and understand short text passages written in English; the listening comprehension test required that students answer 51 multiple-choice questions in real-time, responding to audio recordings of short English passages. Following from the definitions of the tests, the within-item multidimensional model was specified according to two dimensions: (1) a general "text comprehension" dimension, representing the abilities required by items on both tests, and (2) an "auditory processing" dimension, specific to items on the listening comprehension test, only. For the within-item multidimensional model, the inter-factor correlation was fixed to zero. A between-item

multidimensional model was also specified where the two dimensions directly reflected the test content as (1) the “reading comprehension” dimension and (2) the “listening comprehension” dimension. For the between-item multidimensional model, the correlation between latent factors was freely estimated. Similar to the Wu and Adams (2006) study, a unidimensional model was also estimated for comparison. The implied Q-matrices for this study are presented in Table 2.3.

Table 2.3:
Q-matrices for Hartig and Höhler (2008)

Items	Uni- dimensional	Between-Item Multidimensional		Within-Item Multidimensional	
		Reading Comp.	Listening Comp.	Text Comp.	Auditory Processing
1 (R)*	1	1	0	1	0
...	1	1	0	1	0
46 (R)	1	1	0	1	0
47 (L)	1	0	1	1	1
...	1	0	1	1	1
91 (L)	1	0	1	1	1

* *R* = Reading; *L* = Listening.

All of the models were estimated according to a generalized 2PL item response model using the Mplus 4.21 software (Muthén & Muthén, 2007), in which factor loadings were constrained to be equal for items loading on the same dimension. While the estimated inter-factor correlation for the between-item multidimensional model was very high ($\rho = 0.91$), the results of this analysis found that both multidimensional models demonstrated better fit than the unidimensional model. The patterns of factor loadings for the two multidimensional models offer differing interpretations of student performance with regards to skills and abilities. While factor loadings for the reading comprehension test on factor 1 (“reading comprehension” or “text comprehension”) were equivalent across models, factor loadings for factor 2 were lower for the within-item

multidimensional model (“auditory processing”) than for the between-item multidimensional model (“listening comprehension”). These results indicate that the within-item multidimensional model decomposes the abilities required for listening comprehension items, providing information about the mixture of skills necessary for successful performance. The between-item multidimensional model, however, simply separates performance according to test content and suggests a high degree of overlap via the inter-factor correlation but does not specifically differentiate skills or abilities.

In the studies above, item responses were modeled according to Q-matrices shown to demonstrate both between- and within-item multidimensionality under a variety of test and sample design characteristics. These Q-matrices are seen to both describe substantive theory and constrain parameter estimation for each of the associated item response models. In each of the studies, the fit statistics were then examined to facilitate discussion and interpretation of the best fitting model and the corresponding Q-matrix.

2.3. Summary of Notational Conventions

Having described the conditions necessary to achieve equivalence between CFA models and MIRT models and the role that Q-matrices can play in both, the following are the notational conventions that will be used in this dissertation:

- i indexes subjects / respondents / persons / examinees; it is removed from most equations in this dissertation for the purpose of clarity;
- j indexes observed / manifest variables, which are scores from test questions / assessment items; in this dissertation, only binary item scores will be modeled,
- k indexes the latent variables / factors / statistical dimensions in a MIRT or CFA model,

- q_{jk} denotes a binary entry in the Q-matrix so that $q_{jk} = 1$ indicates that item j loads on / requires / measures attribute / latent factor / dimension k for a successful response, and $q_{jk} = 0$ indicates that item j does not load on / require / measure attribute / latent factor / dimension k .
- θ_k , denotes the k^{th} continuous latent variable in a MIRT or CFA model,
- *item difficulty* is represented by either b_j and d_j in unidimensional IRT models, MDIFF in MIRT models, and by τ_j , the threshold parameter, in CFA models.
- *item discrimination* is represented by a_{jk} in unidimensional IRT models, $MDISC_k$ in MIRT models, and λ_{jk} in CFA models.

2.4. Properties of Model-Fit Indices

Given a set of j -observed variables, the covariance structure hypothesized in CFA is $\Sigma_0 = \Sigma(\omega)$, where $\Sigma(\omega)$ is a $j \times j$ covariance matrix of the observed variables or items in the population, $\Sigma(\omega)$ is the model-implied covariance matrix, and ω is a vector of free or estimated parameters in the model. Sample estimates, $\hat{\omega}$, are calculated that minimize the discrepancy between the model implied covariance matrix, $\Sigma(\hat{\omega})$, and the observed covariance matrix, S , according to the discrepancy function $\hat{F}[S, \Sigma(\hat{\omega})]$. The larger the discrepancy, the greater the value of \hat{F} ; therefore, model parameters are estimated such that they minimize the value of the discrepancy function.

There are many estimators of the minimum fit function (F_{min}) but the weighted least squares mean- and variance-adjusted estimator (WLSMV; Muthén & Muthén, 1998-2001; Muthén, Du Toit. & Spisic, 1997) has been shown to be most appropriate for estimating CFA models when the observed variables are dichotomous. Similar to normal

theory estimators, the WLSMV requires the calculation of a full weight matrix, which is the asymptotic covariance matrix that contains tetrachoric correlation estimates when binary responses are modeled; however, only the diagonal of this weight matrix is used to calculate factor model parameter estimates. Subsequent to parameter estimation, the full asymptotic covariance matrix is again employed to calculate the goodness-of-fit χ^2 which then has a mean and variance adjustment factor applied to account for the categorical nature of the data (Muthén, Du Toit, & Spisic, 1997).

The use of normal-theory estimators, such as *maximum likelihood* (ML) and *generalized least squares* (GLS) methods have been shown to produce inflated goodness-of-fit chi-square estimates when modeling categorical data (Bentler & Dudgeon, 1996; Bollen, 1989; Finney & DiStefano, 2006). As implemented in the Mplus software package, version 5 (Muthén & Muthén, 1998-2007), results have shown that the WLSMV yields acceptable Type-I error rates and parameter estimate bias when three-dimensional models were estimated with 12 observed variables and sample sizes ranging greater than 200 (Muthén et al., 1997).

One of the most popular methods for evaluating model fit is the goodness-of-fit χ^2 statistic (Hu & Bentler, 1999), which assesses the magnitude of discrepancy between the estimated and predicted covariance matrices as follows:

$$\chi^2 = F_{min}(N - 1)$$

where N denotes the sample size. It follows χ^2 distribution when the model is correctly specified, with an expected value equal to the degrees of freedom (df) and variance of

$2df^3$. A significant χ^2 value may reflect model misspecification, including violations of some of the underlying assumptions (Hu & Bentler, 1998).

While the χ^2 statistic features prominently in the adjudication of model fit (Gierl & Mulvenon, 1995); a variety of other model fit indexes exist, having been developed to overcome the shortcomings of this statistic, specifically its sensitivity to sample size (Fan, Thompson, & Wang, 1999) and the violation of distributional assumptions. These statistics can be classified as *incremental*, *absolute*, and *parsimony-adjusted* fit indexes (Bandalos & Finney, 2010; Bollen, 1989; Gerbing & Anderson, 1993; Hu & Bentler, 1995; Marsh, Balla, & McDonald, 1988; Tanaka, 1993).

Incremental or baseline fit indices (Curran, Bollen, Chen, Paxton, & Kirby, 2003) calculate the improvement in model fit offered by the hypothesized, estimated, model in comparison with a more restricted, nested, baseline model. Typically, this null model considers all observed variables to be uncorrelated (Bandalos & Finney, 2010; Bentler & Bonett, 1980). Incremental fit indices, however, are excluded from this dissertation as they have been shown to demonstrate undesirable sensitivity to factors such as sample size and number of observed variables while being minimally sensitive to Q-matrix misspecification, the model characteristic of interest in this dissertation (Beauducel, & Wittmann, 2005; Fan & Sivo, 2005, 2007; Fan, Thompson, & Wang, 1999; Hartig & Höhler, 2008; Jackson, 2007; Janssen & De Boeck, 1999; Marsh, Hau, & Wen, 2004; Sivo, Fan, Witta, & Willse, 2006; Thurber, Shinn, & Smolkowski, 2002; Wolfe, Hickey, & Kindfield, 2009).

³ Model misspecification, however, results in a non-central χ^2 distribution with an expectation of $df + \lambda$ (the non-centrality parameter) and variance of $2df + 4\lambda$ (Curran et al., 2003; Steiger, Shaprio, and Browne, 1985).

Absolute fit indices are another category of goodness-of-fit statistics, which assess how well an *a priori* model reproduces the sample data. No reference model is used to assess the amount of increment in model fit, but an implicit or explicit comparison may be made to a saturated model that exactly reproduces the sample covariance matrix. Included in this category is the classic goodness-of-fit χ^2 statistic, described above, as well as alternatives to this model fit index such as McDonald's GFI index (McDonald, 1989), and the standardized root-mean-square residual (SRMR; Jöreskog & Sörbom, 1981; Steiger, 1989). Though Hu and Bentler (1998) included the root mean square of error approximation (RMSEA) in this category, it is more appropriately classified as a parsimony-adjusted fit index (Browne & Cudeck, 1993; Steiger & Lind, 1980).

Similar to absolute fit indices are the parsimony-adjusted indices which also measure the discrepancy between observed and model-implied covariances, but also incorporate some type of *penalty* adjusting for degrees of freedom or model complexity. Therefore, these indices describe the amount of increment in model fit relative to the number of parameters required to obtain this increase in model fit. These indices include the Parsimony Goodness of Fit Index (PGFI) and Parsimonious Normed Fit Index (PNFI), which were developed by Mulaik et al. (1989) to overcome issues with the GFI and NFI incremental fit indices and include a parsimony ratio, $\frac{df}{df_0}$, in which the degrees of freedom for the hypothesized model, df , are divided by the degrees of freedom for the null model, df_0 . As it also accounts for model complexity, the RMSEA model-fit index is included in this category and described in detail later.

Model, or global, fit indices considered in this study were selected from the families of absolute and parsimony-adjusted fit indices based on three primary criteria. First, they had to have been rather frequently investigated by researchers so that a strong empirical research base was already available upon which this dissertation work sought to expand. Second, they had to have shown sensitivity to measurement model misspecification in previous work. Third, they had to have been shown to be sufficiently robust to test design conditions in previous work.

Utilizing these criteria, the selected indices for this dissertation were the χ^2/df ratio, an absolute fit index adjusted for model complexity, and the *root mean square error of approximation* (RMSEA). Both of these indices are available in most CFA and SEM software packages.

This dissertation also investigates the *generalized dimensionality discrepancy measure* (GDDM; Levy & Svetina, 2010), which was developed in application under the MIRT framework. Given the equivalence between CFA and MIRT models established previously, all three of these statistics can be employed in the evaluation of model-fit when the appropriate modeling assumptions have been met. The following sections describe the structure of and prior research on these indices in more detail.

2.4.1 The χ^2/df Model-Fit Index

The χ^2/df ratio model-fit statistic is simply a rescaling of the goodness-of-fit χ^2 index described earlier according to the model degrees of freedom which has been recommended as appropriate for evaluating models under conditions of model misspecification (Beauducel & Wittman, 2005; Jackson, 2007; Marsh, Hau, & Wen, 2004).

In simulation studies, the χ^2/df has demonstrated values that increase with sample size and model complexity increased when misspecified models are estimated (Beauducel & Wittman, 2005; Jackson, 2007; Marsh, Hau, & Wen, 2004), suggesting that this statistic becomes more powerful under these conditions. It has also been shown to demonstrate stable nominal Type-I error rates (Marsh, Hau, & Wen, 2004) and generally outperformed all other fit indices in correctly rejecting misspecified models as one of the model-fit indices most sensitive to model misspecification (Jackson, 2007; Marsh, Hau, & Wen, 2004). Additionally, Wolfe, Hickey, and Kindfield (2009) found that the χ^2/df model-fit index was able to distinguish between competing MIRT models of two and three dimensions when applied to real-world data describing student performance on a test of introductory genetics.

2.4.2 The RMSEA Index

The RMSEA index (Browne and Cudeck, 1993; Steiger, 2000; Steiger and Lind, 1980) is a parsimony-adjusted model fit index which is based on a non-central goodness-of-fit (GOF) χ^2 . The sample estimate of RMSEA (\hat{e}) is calculated as:

$$\hat{e} = \sqrt{\hat{F}_0 / df} = \sqrt{\frac{GOF \chi^2 - df}{df(N-1)}} = \sqrt{\frac{\hat{\lambda}}{df(N-1)}}$$

where \hat{F}_0 is the sample estimate of the error of approximation (Browne and Cudeck, 1993) or the degree of misfit between the population covariance matrix (Σ_0) and the model-implied population covariance matrix (Σ_0^*) – which is estimated as the discrepancy function $F_0 = (\Sigma_0, \Sigma_0^*)$ according to the specified estimation procedure. Given that the degrees of freedom can exceed the GOF χ^2 , the minimum of this

denominator is set to zero and the sample estimate of $\hat{\epsilon}$ ranges from zero to infinity, where zero indicates perfect fit and larger values indicate worse fit. The degrees of freedom also indicate the number of dimensions by which the data are free to differ from a model with estimated parameters; the RMSEA is a measure of the average lack of fit per the degrees of freedom or potential lack of fit (Heene, Hilbert, Draxler, Ziegler, and Bühner, 2011).

In the seminal paper by Hu and Bentler (1998), a large body of literature on model fit was used to inform the design of a simulation study for the purpose of evaluating the performance of model-fit indices, including the RMSEA. This study has informed a great deal of subsequent research into the performance of the RMSEA which has been shown to demonstrate appropriate sensitivity to model misspecification while also maintaining minimal sensitivity to other factors and has been specifically recommended for use in detecting measurement model misspecification (Beauducel & Wittman, 2005; Curran et al., 2003; Fan & Sivo, 2005, 2007; Fan, Thompson, & Wang, 1999; Jackson, 2007; Sivo, Fan, Witta, & Willse, 2006).

Examination of the performance of the RMSEA has found that it demonstrates minimal-to-modest sensitivity to various factors, defined as the proportion of variance in the RMSEA attributed to the specific source. Factors to which the RMSEA has been shown to be minimally sensitive include sample size, the distributional form of observed continuous responses, and estimation method (i.e., Maximum Likelihood, Generalized Least Squares, or Asymptotic Distribution Free estimation) (Hu & Bentler, 1998; Beauducel & Wittman, 2005; Curran et al., 2003; Fan & Sivo, 2005, 2007; Fan, Thompson, & Wang, 1999; Sivo, Fan, Witte, & Willse, 2006). Additionally, values of the

RMSEA have been shown to increase with number of latent factors under true and misspecified model estimation (Beauducel & Wittman, 2005; Fan & Sivo, 2007). When models were misspecified, the RMSEA has shown sensitivity to such model misspecification, typically as a result of under-factoring (Hu & Bentler, 1998; Fan & Sivo, 2005; Fan, Thompson, & Wang, 1999), corresponding to large discrepancies between values resulting from correctly estimated models in comparison to those estimated from the misspecified models (Sivo, Fan, Witte, & Willse, 2006). Further, the RMSEA demonstrates acceptable power rates when rejecting misspecified models (Beauducel & Wittman, 2005). A final important consideration is that the RMSEA has shown little systematic bias and random variation in simulation studies for sample sizes of $n = 200$ or greater (Curran et al., 2003; Fan, Thompson, & Wang, 1999). All of these results suggest the RMSEA as an appropriate model-fit index for inclusion in this dissertation.

2.4.3 The Generalized Dimensionality Discrepancy Measure

The *generalized dimensionality discrepancy measure* (GDDM) is a new model-fit statistic that was developed original under a posterior predictive model-checking (PPMC) framework for MIRT (Levy & Svetina, 2010). Under a correctly specified model, responses to items for a given person are locally independent if

$$P(\mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\omega}) = \prod_{j=1}^J P(x_j | \boldsymbol{\theta}, \boldsymbol{\omega}_j) \text{ where } P(x_j | \boldsymbol{\theta}, \boldsymbol{\omega}_j) \text{ is the item response function for item } j$$

given student ability $\boldsymbol{\theta}$ over k -dimensions and $\boldsymbol{\omega}$ is the collection of item-specific parameters accounting for the presumed MIRT model. Violations of this assumption are a

result of model-data misfit and produce biased item parameter estimates, test statistics, and student ability estimates (Zenisky, Hambleton, Sireci, 2002).

The GDDM is essentially the mean of the absolute squared differences between observed and expected responses computed over all unique item pairs:

$$GDDM = \frac{\sum_{j \neq j'}^J \left| N^{-1} \sum_{i=1}^N (X_{ij} - E(X_{ij} | \theta_i, \omega_j)) (X_{ij'} - E(X_{ij'} | \theta_i, \omega_{j'})) \right|}{J(J-1)}.$$

Values of this statistic range from zero, indicating no conditional covariance between all items on a test, to infinity with larger values indicating greater dependence. Large GDDM values, therefore, indicate poor model fit.

In a Monte Carlo simulation by Levy and Svetina (2010), the GDDM was found to perform at nominal levels in identifying misfit, violations of local independence, when applied to two- and three-dimensional 2PNO MIRT models. Additionally, the GDDM was used to examine responses to the 1996 *National Assessment of Educational Progress* (NAEP) in science according to a three-dimensional, 3PNO MIRT model. Applied to the actual responses of 1,020 examinees to 16 items, the GDDM coupled with item-level information provided by the MBC was successfully used to diagnose overall test performance and identify misfitting items which are candidates for subsequent review.

This section described the properties of the χ^2/df , RMSEA, and GDDM model-fit indices, including the mathematical foundations and brief reviews of previous research. These fit indices have been selected for inclusion in the current dissertation since they have demonstrated desirable performance for the purpose of detecting CFA or MIRT model misspecification while being minimally sensitive to other factors. In the following section, item-fit indices are similarly considered.

2.5. Properties of Item Fit Indices

Item fit analysis is concerned with the assessment of model-data fit at the level of individual score variables, rather than at the aggregate level that the model-fit statistics represent. Under the CFA framework the two most commonly used local fit indices are the *Modification Index* (MI; Sörbom, 1989) and the *Wald Test* (Buse, 1982). Within the IRT framework there has been comparatively little research on item-fit for MIRT models even though the $S\text{-}\chi^2$ statistic has been shown to be a potentially promising candidate based on preliminary research (Zhang & Stone, 2008; see also Li & Rupp, 2012). As with the model-fit indices described previously, equivalence between the CFA and MIRT models allows these item-fit indices to be applied to a wide variety of latent variable models.

2.5.1 The $S\text{-}\chi^2$ Statistic

Though numerous unidimensional IRT item-fit indices have been proposed, very little research on item fit indices under a MIRT framework has been conducted. One statistic that has been investigated is the $S\text{-}\chi^2$ statistic proposed by Orlando and Thissen (2000) which has been subsequently adapted for application to MIRT models (Zhang & Stone, 2008). This statistic is a desirable candidate because (1) it employs total score and does not rely on ability estimation, (2) the statistic is a function of observed proportions making the null distribution easy to describe, and (3) the contingency table required to compute the statistic is manageable in size which has the additional effect of limiting the potential for sparse data structures.

The performance of the $S\text{-}\chi^2$ statistic within a MIRT framework was evaluated by Zhang and Stone (2008) using a Monte Carlo simulation design examining Type-I error

rates and power in detecting a misfitting item that either violated monotonicity or ignored guessing. When data was generated under simple structure, the Type-I error rates were appropriate for all other conditions. When the data were generated according to complex structure, however, Type-I error rates were inflated when the dimensions were highly correlated and when the sample size was large.

Across conditions, the power to detect violation of monotonicity increased across sample sizes to nearly 100%, and increased with inter-factor correlation, demonstrating the highest power rates by correlations of 0.6 which persist for stronger correlations. Power to detect item misfit due to ignoring a guessing effect was low to moderate, increasing with sample size and inter-factor correlation. The results of this study show the $S\text{-}\chi^2$ statistic to be a viable option for assessing item fit under a MIRT framework, though it results in “liberal rejection of model-fitting items” (Zhang & Stone, 2008, p. 193) when the test structure is complex.

2.5.2 The Modification Index

Modification indices (MI; Sörbom, 1989) are a function of the first order derivatives of the fitting function evaluated at each fixed parameter or factor loading and are scaled to a χ^2 metric (Kaplan, 1991). MI values reflect the approximate decrease in the overall model χ^2 if the current parameter were freely estimated and are, therefore, analogous to the χ^2 difference test or likelihood ratio between two nested models. The use of the MI has been shown to facilitate revision of misspecified models when the revision is theoretically justifiable and substantively interpretable (Jöreskog, 1993; Kaplan, 1989, 1990; MacCallum, 1986; Silvia & MacCallum, 1988).

A Monte Carlo simulation study was conducted by Hutchinson (1998) to examine the stability of the results of an automated *specification search* or successive sequential revision according to significant MI values when applied to two- and four-factor CFA models estimated according to four levels of severity of misspecification. Her results found that recovery of the population model as a result of MI-based model revision improved as sample size increased from 200 to 1200 and worsened as the severity of misspecification increased, defined according to number and magnitude of factor loadings fixed to zero. When misspecification was slight, stability was achieved and the population model recovered at least 90% of the time at $n = 800$ for the two-factor model and $n = 1200$ for the four-factor model; under severe misspecification, the four-factor model better recovered the population model and achieved 90% recovery at $n = 1200$. Overall, the study suggests that MI is useful though sensitive to sample size, model complexity, and the magnitude of omitted factor loadings.

2.5.3 The Wald Test

The *Wald Test* (Buse, 1982) is a univariate χ^2 typically presented as the square of the normal z -value for each freely estimated parameter, and can be thought of as complementary to the MI as it indicates whether a freely estimated parameter should be fixed or set to zero. This local-fit statistic has been shown to be asymptotically equivalent to the likelihood-ratio test between two nested models (Buse, 1982; Kaplan, 1989).

Even though this index has been rarely studied empirically, a Monte Carlo simulation by Chou and Bentler (2002) examined the performance of the Wald Test in *backward searches* on misspecified structural parameters in two different SEM models. When the saturated model contained fewer misspecifications, the Wald Test was able to

correctly reject parameters in 75% of the replications while incorrectly rejecting true nonzero parameters 12% of the time; success rates improved when candidate parameters were limited according to theoretical justification. For the saturated model that contained a greater number of misspecified parameters, misspecified parameters were rejected greater than 65% of the time and true nonzero parameters were rarely rejected; performance increased to greater than 95% when selection was limited by theoretical justification.

2.6. Summary

Prior to description of the simulation study conducted in this dissertation, this chapter described the necessary mathematical conditions establishing equivalence between CFA and MIRT models. Making the assumptions that unobserved continuous response are manifest as dichotomous item responses, that errors are normally and independently distributed, and that latent factors follow a multivariate normal distribution with unit variance, parameters resulting from CFA and MIRT models are seen to be equivalent through known transformations.

Description of the Q-matrix, a structure that both operationalizes substantive theory as well as serving to constrain model parameters, provides additional information necessary to understand the correspondence between CFA and MIRT models as well as providing a clear device within which model misspecification may be expressed. In the CFA context, the Q-matrix may represent either simple or complex structure and defines the pattern of fixed and freely-estimated factor loadings. In the MIRT context, the Q-matrix represents between- or within-item multidimensionality as binary elements indicating fixed or freely-estimated item discrimination parameters.

Further, owing to the equivalence between CFA and MIRT models, model- and item-fit statistics typically available separately for these two models may be employed together in the evaluation of model fit under true, correctly, estimated models and misspecified models. Previous research having described or demonstrated appropriate and desirable qualities in detecting model misspecification under CFA or MIRT models while being minimally sensitive to other factors, the model-fit indices included in the subsequent simulation study are the χ^2/df , RMSEA, and GDDM and the item-fit indices are the $S-\chi^2$, Modification Index, and Wald Test.

Chapter 3

Methods

3.1.Objective

This study seeks to examine the performance of various global and local fit indices under Multidimensional Item Response Theory (MIRT) and Confirmatory Factor Analysis (CFA) frameworks according to different test design and respondent population conditions. Specifically, factors that include sample size, test length (number of items), model complexity (simple- or complex-structure), model dimensionality (number of latent factors), inter-factor correlation, and item type (jointly defined by difficulty and discrimination) will be manipulated within a simulation study for the purpose of answering the following research questions.

- 1) In terms of baseline performance under correct model specification:
 - a) How well are key parameters (e.g., item difficulties, item discriminations, inter-factor correlations, person estimates) recovered under various simulation conditions, as indicated by average bias and root mean-squared error?
 - b) How do cut-off points associated with different significance levels (0.10, 0.05, 0.01) resulting from the empirical sampling distributions for each fit statistic align with those of the theoretical sampling distributions under different simulation conditions?
 - c) What proportion of variance in the empirically-derived cut-off values in each fit index is accounted for by each of the simulation conditions?
- 2) In terms of performance under model misspecification – specifically Q-matrix misspecification:

- a) How is item parameter recovery affected by Q-matrix misspecification under different simulation conditions?
 - b) What is the power of different fit indices to detect Q-matrix misspecification using the empirically-derived cut-off values as suggested under correct model specification?
 - c) What proportion of variance of power values of the different fit indices is accounted for by each of the simulation conditions?
- 3) How can the findings from the simulation studies be used to evaluate and revise potentially misspecified Q-matrices for real data sets when the data-collection design conditions are similar to the simulation design conditions?

3.1.1 Simulation Conditions

In this section, the specific conditions employed during the data generation phase of this dissertation are described. Sample size, test length, model dimensionality and complexity, and Q-matrix structure are often directly manipulated by or under the control of researchers whereas item characteristics such as difficulty and discrimination as well as correlational dimensions are model parameters that are estimated and not directly controllable. The full simulation design is summarized in Table 3.1.

Table 3.1:
Simulation Design Summary

Purpose	Condition	Levels	Description
Generation	Model Dimensionality	2	Low (2 latent factors); Moderate (3)
	Test length	3	Short (12 items); Moderate (24 items); Long (36 items)
	Respondent sample size	2	Small (n = 250), Large (n = 1,000)
	Model Complexity	2	Simple-structure; Complex-structure
	Inter-factor correlation	3	Weak (r = 0.25); Moderate (r = 0.50); Strong (r = 0.75)
	Item type (disc. & diff.)*	6	HH; HM; HL; MH; MM; ML
	Total	432	
Estimation	Framework	2**	CFA; MIRT
	(Mis)specification	3	True model; Moderate (17%); Severe (33%)
	Total	6	
Total		1296***	

* Item type is denoted according to discrimination and item difficulty discrepancy from the population mean ("difficulty"). *H* = high discrimination or difficulty; *M* = moderate discrimination or difficulty; and *L* = low difficulty only.

* Though two estimation frameworks are specified, models are only estimated once given the equivalence of MIRT and CFA under the conditions specified for this study.

** The conditions represent a fully-crossed simulation design.

3.1.1.1 Model Dimensionality

The number of latent variables assigned to subjects, examinees, or students in this study will represent two or three abilities, attributes, or dimensions. These latent factors, θ_k , will be generated as $\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $k \times k$ covariance matrix described according to the levels of the inter-factor correlation condition. Though two or three factors seem small, previous studies examining global or local fit under CFA or MIRT frameworks typically considered few latent factors (e.g., Fan & Sivo, 2005, 2007; Finch, 2010, 2011; Ximénez, 2009). Three latent factors was the median number of first-order

factors reported in reviews of CFA studies by Baumgartner and Homburg (2006) and Jackson, Gillaspy, and Purc-Stephenson (2009).

3.1.1.2 Test Length

Describing the number of observed variables or test items present on an instrument, the current study specifies a short test length (12 items), a medium test length (24 items) and a long test length (36 items). These lengths are chosen to reflect prototypical educational assessment conditions in that shorter tests are typically applied in classroom settings by teachers while longer tests are common in large-scale, high-stakes assessment situations. Moreover, these lengths ensure equal numbers of items per latent factor for each of the latent factor conditions. For example, a short test of 12 items yields 6 items per factor under the 2-factor model and 4 items per factor under the 3-factor model; similarly, a longer test of 36 items yields 18 items per factor under the 2-factor model and 12 items per factor under the 3-factor model. This follows the practice of previous research (de la Torre, 2008; Henson & Templin, 2006) and ensures that the same numbers of pieces of statistical information are available on each latent factor for estimating respondent parameters.

The number of observed variables or items considered in previous studies on local or global fit vary widely; the minimum number of items per dimension was four while the maximum was sixty and the minimum total number of items was four and the maximum was 97. The median number of observed variables reported in the review by Baumgartner and Homburg (2006) was 11 while that reported by Jackson, Gillaspy, and Purc-Stephenson (2009) was 17 with both reporting ranges lower than 10 and greater than 20. Finch (2010, 2011) has shown that item parameter recovery is largely unaffected

by test length, which is important to consider as item parameters are instrumental in the calculation of local fit indices like the $S\text{-}\chi^2$. This is confirmed by findings from Orlando and Thissen (2003) who showed that the $S\text{-}\chi^2$ demonstrated favorable detection rates for misfitting items when the tests were composed of more than 10 items.

3.1.1.3 Sample Size

Manipulating the number of observations, small ($n = 250$) and large ($n = 1000$) sample sizes will be employed in the current study. Based on the range of sample sizes reported across studies under the CFA framework (ranging 30 to 5,000; Beauducel, & Wittmann, 2005; Fan & Sivo, 2005; Fan & Sivo, 2007; Fan, Thompson, & Wang, 1999; Jackson, 2007; Marsh, Hau, & Wen, 2004; Sivo, Fan, Witta, & Willse, 2006; Thurber, Shinn, & Smolkowski, 2002) and the MIRT framework (200 to 10,000; Hartig & Höhler, 2008; Janssen & De Boeck, 1999; Wolfe, Hickey, Kindfield, 2009) a sample size of 250 represents an acceptable lower bound across CFA and MIRT studies while approximately one-quarter of the CFA studies employed sample sizes of 1,000 or greater. Reviews of studies applying CFA models in marketing and consumer research (Baumgartner & Homburg, 2006) found that sample sizes ranged $n = 143$ to $n = 305$, suggesting the small sample size; much larger sample sizes were found in reviews of CFA applications in the field of social work ($n = 120$ to $6,424$; Guo, Perron, & Gillespie, 2009) and in journals of applied, counseling, and personality psychology ($n = 58$ to $46,133$; Jackson, Gillaspay, & Purc-Stephenson, 2009) While measures of model fit, especially the RMSEA, have been found to be largely insensitive to sample size (Fan & Sivo, 2005, 2007; Ximénez, 2009), the $S\text{-}\chi^2$ index performs marginally well at sample sizes of 500 and favorably at sample

sizes of 1000. Therefore, the sample sizes used in this study should allow for an appropriate detection of the degree of sensitivity of these indices.

3.1.1.4 Model Complexity

A key design characteristic of an instrument is the number of latent factors associated with each item. Recall that the characteristic of item multidimensionality (Adams, Wilson, & Wang, 1997) is represented via row vectors in the Q-matrix whereby items associated with a single latent factor are referred to as *between-item multidimensional* and items associated with multiple latent factors are referred to as demonstrating *within-item multidimensionality*. In CFA terminology, a simple-structure model is composed entirely of items demonstrating between-item multidimensionality while a complex-structure model is comprised of at least one item demonstrating within-item multidimensionality. In maintaining the MIRT and CFA terminology, the dimensionality of items and models is differentiated by referring to the former as between- or within-item multidimensional and referring to the latter as simple- and complex-structure.

A separate Q-matrix following simple-structure, comprised solely of items demonstrating between-item multidimensionality, is specified for each combination of the levels of the Test Length (i.e., 12 items, 24, and 36) and the Model Size (i.e., 2 latent factors or 3), resulting in 6 between-item multidimensional Q-matrices which are presented in the Appendix. Each of these Q-matrices is constructed such that the k marginal column proportions – the number of items associated with each of the k latent factors – is equal, providing a degree of measurement consistency since the generating

item difficulty and discrimination (or factor loading) values are similar within Item Type conditions.

Q-matrices following complex-structure, containing items demonstrating within-item multidimensionality, are constructed using the simple-structure Q-matrices as a starting point. For this condition, one-third of the j items in the two latent factor condition and one-quarter of the j items in the three latent factor condition are defined in the respective Q-matrix as within-item multidimensional and strategically associated with a second latent factor, $q_{jk} = 1$, such that the equality of the marginal column proportions was maintained. The remaining items in each Q-matrix were left specified as between-item multidimensional. These Q-matrices are also presented in the Appendix.

In simulation studies conducted by Fan and Sivo (2005, 2007) and Hu and Bentler (1998, 1999) model fit was found to be better for misspecified models when the generating model followed simple-structure and the estimating models followed complex-structure; model fit was comparatively worse for those models generated as complex-structure and estimated as simple-structure. At the local, or item, fit level, Zhang and Stone (2008) found that under conditions of between-item multidimensionality, Type-I error rates for detecting misfitting items approached the nominal level while Type-I error rates were inflated under within-item multidimensionality and especially as test length and inter-factor correlation increased. Finch (2011) notes that within-item multidimensionality produces overestimates of MIRT discrimination parameters and underestimates of difficulty parameters, thereby affecting measures of item fit when the model is correctly specified.

3.1.1.5 Inter-Factor Correlation

The two or three latent factors assigned to each examinee in this study are specified as correlated to a certain degree. The current study considers weak ($r = 0.25$), moderate ($r = 0.50$), and strong ($r = 0.75$) inter-factor correlations, equal for all pairs of factors. Studies emulating the results of Hu and Bentler (1998) employed inter-factor correlations of 0.3 to 0.5 (Fan & Sivo, 2005, 2007); inter-factor correlations included in the study by Ximénez (2009) ranged 0.3 to 0.9. The studies by Finch (2010, 2011) found that as inter-factor correlation increased from 0.0 through 0.8 so too did the bias in item parameter estimates, suggesting sensitivity of local fit indices to such dependencies.

3.1.1.6 Item Types

Further unobservable characteristics of instruments and variables, though controllable in a simulation study, are the item difficulty and discrimination parameters. In this dissertation, discrimination and difficulty are fully-crossed and specified jointly according to six item types:

- High discrimination / high difficulty (HH);
- High discrimination / moderate difficulty (HM);
- High discrimination / low difficulty (HL);
- Moderate discrimination / high difficulty (MH);
- Moderate discrimination / moderate difficulty (MM); and
- Moderate discrimination / low difficulty (ML).

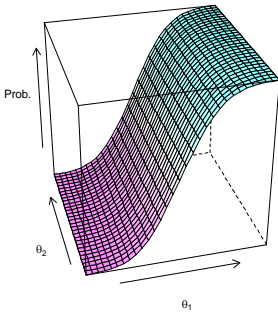
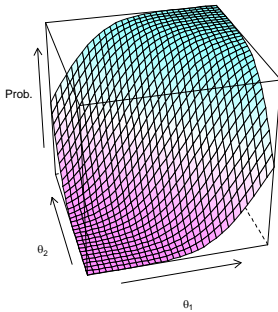
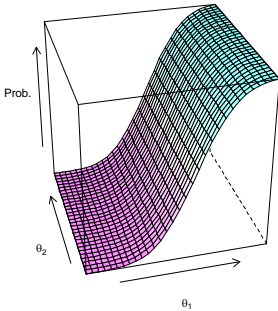
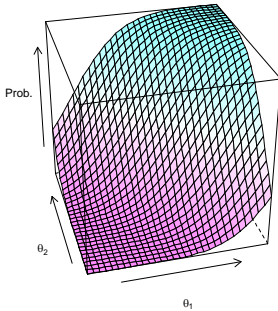
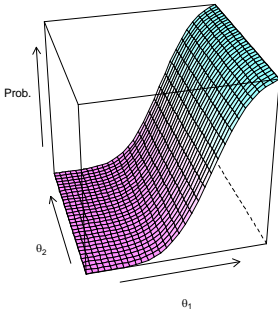
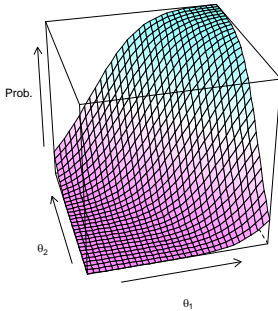
Item difficulty and discrimination values vary across Model Complexity, Model Size, Test Length, and Item Type conditions, resulting in 72 parameter sets which are constant across all other data generation conditions; the exact values are presented along

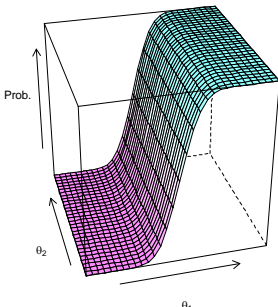
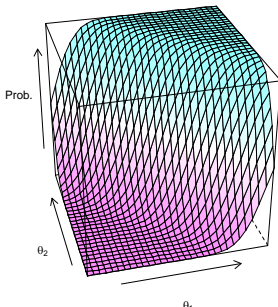
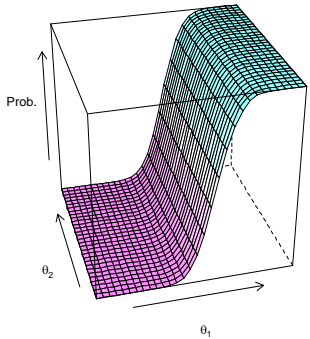
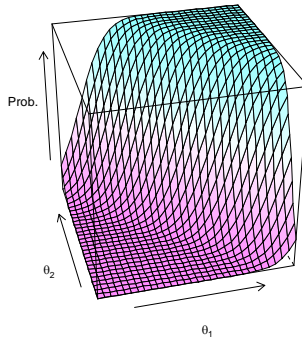
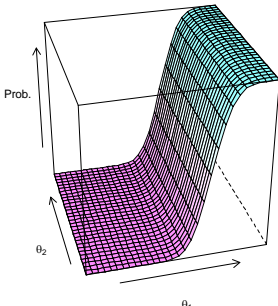
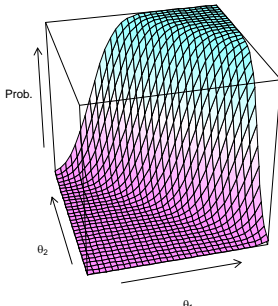
with the Q-matrices in the Appendix. Multidimensional item discrimination (MDISC) values for this study are strategically distributed as a $j \times 1$ vector across items with range = $[+0.9, +1.1]$ and mean = $+1.0$ for the moderate discrimination conditions while range = $[+1.4, +1.6]$ and mean = $+1.5$ for the high discrimination conditions. Higher discrimination values serve to differentiate clearly among examinees while the moderate discrimination condition approximates the Rasch model (Embretson & Reise, 2000; Rasch, 1960/1980), frequently applied in the analysis of assessment data.

Multidimensional item difficulty (MDIFF) values in this study are specified according to the degree of discrepancy between the distribution of item difficulty parameters and the distribution of the generating ability parameters, θ ; the suffix “difficulty” is retained instead of “discrepancy” to facilitate later discussion and labeling. Low difficulty items represent low discrepancy and are well-targeted to the ability distribution in the population; moderate difficulty items represent moderate discrepancy and the distribution is, therefore, slightly shifted away from the examinee ability distribution; lastly, high difficulty items represent high discrepancy and the distribution of item difficulty values is severely shifted away from the distribution of examinee ability. Degree of discrepancy in the current study is manipulated by shifting the distribution of MDIFF parameters from an approximately normal distribution under the low difficulty conditions to a strongly-negatively skewed distribution under the high difficulty conditions; mean difficulty increases with discrepancy, resulting in fewer correct responses by the simulated examinees. Since previous research has shown that item fit is not sensitive to item difficulty (Dodeen, 2004; Reise, 1990), conditions of increasing discrepancy were selected over conditions representing easy or difficult items,

since the latter would likely present redundant fit information. Similar to the MDISC values, MDIFF values are also strategically distributed across items as a $j \times 1$ vector, where mean = 1.0 (approximately) for high difficulty items, mean = 0.50 (approximately) for moderate difficulty items, and mean = 0.0 for low difficulty. A half-logit increase in MDIFF across conditions approximates the difficulty increase between grades described by Kolen and Tong (2010). Further, MDIFF values for all conditions in the current study are defined by range = [-2.0, +2.0] thus ensuring that items represent and provide information across the range of ability of approximately 95% of the simulated examinees. Table 3.2 presents surface plots for prototypical items of each type, as both between- and within-item multidimensional. Additionally, Figure 3.1 presents the kernel-smoothed density plots of the distribution MDIFF values by number of items and difficulty together with the means and inter-quartile ranges.

Table 3.2:
MIRT Surface Plots for Each Item Type

Item Type	MDIFF / MDISC	Between-Item Multidimensional	Within-Item Multidimensional
LM	0.0 / 1.0		
MM	0.5 / 1.0		
HM	1.0 / 1.0		

Item Type	MDIFF / MDISC	Between-Item Multidimensional	Within-Item Multidimensional
LH	0.0 / 2.0		
MH	0.5 / 2.0		
HH	1.0 / 2.0		

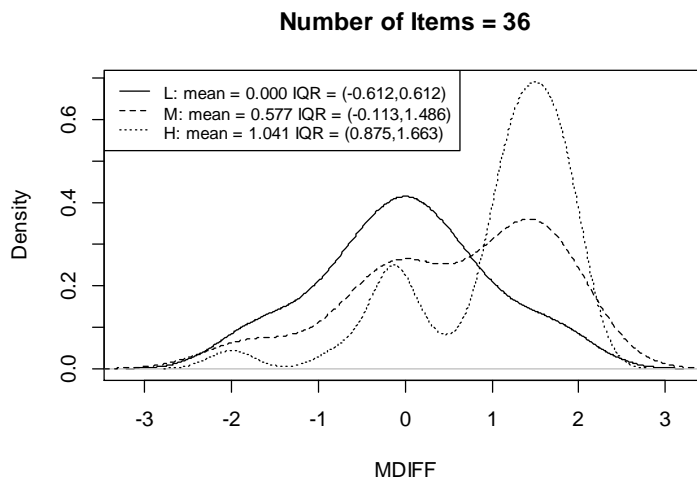
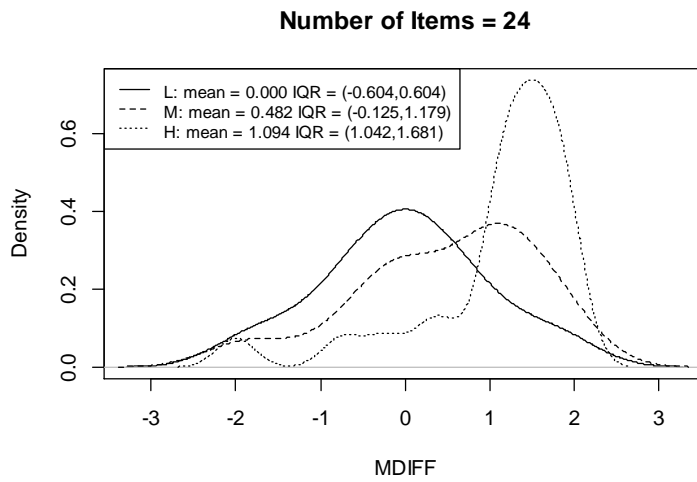
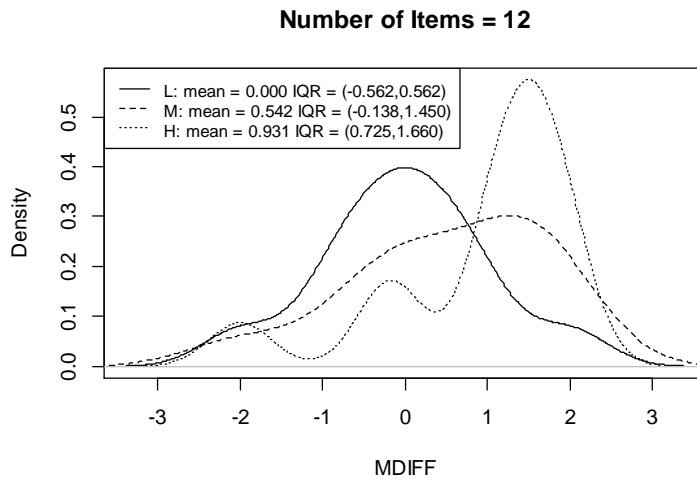


Figure 3.1. Kernel-smoothed density plots of the distributions of MDIFF values by Test Length and difficulty.

These item types do not cover the full range of difficulty and discrimination parameter combinations but reflect a selection similar to the values employed in MIRT studies such as Finch (2011) and Zhang and Stone (2008). Item discrimination values for simple-structure models in the study by Finch (2011) were randomly generated as $\mathbf{a}_1 \sim N(0.9657, 0.3161)$ and constrained within 0.7 and 2.0. For complex-structure models, secondary dimensions were assigned by specifying additional randomly-generated discrimination values $\mathbf{a}_2 \sim N(0.35, 0.15)$ with a minimum of 0.10 and a maximum of 0.60. Item difficulty was randomly generated as $\mathbf{b} \sim N(0,1)$. Zhang and Stone (2008) randomly generated the discrimination values for the first factor in a MIRT model as $\mathbf{a}_1 \sim U[0.4, 2.0]$ then determined the discrimination values for the second factor by randomly sampling the composite angle as $\gamma \sim U[0,20]$ for simple-structure models or $\gamma \sim U[20,45]$ and calculating the remaining discrimination values for each j item in closed form as $a_{j,2} = \sqrt{1/2[a_{j,1}/\cos(\gamma_j)]}$ (Reckase, 2009). The multidimensional difficulty (MDIFF) and discrimination (MDISC) values corresponding to the ranges employed in the above studies are presented in Table 3.3.

Table 3.3:
Summary of MIRT Item Parameters

Study	Model Complexity	MDIFF		MDISC	
		Min	Max	Min	Max
Zhang & Stone (2008)	Simple	-5.000	5.000	0.600	2.540
	Complex	-3.721	3.721	0.806	2.430
Finch (2011)	Simple	-2.121	2.121	0.990	2.828
	Complex	-2.121	2.121	0.141	0.849

Studies examining the impact of model misspecification under the CFA framework have typically emulated the approach and specifications of Hu and Bentler (1998) whereby item discrimination values range 0.98 to 1.33 (e.g., Curran et al., 2003;

Fan & Sivo, 2005, 2007); Ximénez (2009) considered a wider range of values, 0.31 to 2.06. More recently, Heene et al. (2011) manipulated a large range (0.3 to 0.9) of factor loadings considering the effect of such parameters on model fit evaluation which reflect a range of item discrimination values (approximately 0.3 to 2.1) greater than typically seen in IRT or MIRT studies. None of these studies, however, manipulated the magnitude of item discrimination as a factor of interest nor did they explicitly consider item difficulty via threshold parameters nor did the studies by Finch (2010, 2011) explicitly manipulate item difficulty or discrimination.

3.1.2 Data Generation.

For each of the 432 data generation cells across the simulation conditions presented in Table 3.1, 1000 replications will be performed under True model misspecification while 250 replications will be performed for the Moderate and Severe misspecification conditions, thus allowing for the examination of distributional properties, the calculation of various descriptive statistics, and the computation of specific analysis-of-variance (ANOVA) models as described below.

For each combination of the simulation factors described, item responses for examinees $i = 1, \dots, I$ (determined by Sample Size) to items $j = 1, \dots, J$ (Test Length) are calculated according examinee ability on each of the k latent factors, θ_{ik} , given the 2-parameter normal-ogive multidimensional item response model (2-PNO MIRT; De Ayala, 2009; Lord, 1952; Reckase, 2009). If the probability of a correct response by examinee i to item j given abilities $\theta_{i1} \dots \theta_{ik}$ is greater than the corresponding value from an $i \times j$ matrix, \mathbf{U} , of random uniform values ranging $[0, 1]$, then a correct item response is generated ($x_{ij} = 1$), otherwise an incorrect item response is generated ($x_{ij} = 0$).

Comparing item response probabilities against \mathbf{U} introduces random error as suggested by Luecht (1996). Data generation was conducted in the R software package (R Development Core Team, 2011). This procedure is summarized in the following outline:

1. *Q-Matrix Generation* – for each level of Model Dimensionality, Test Length, and Item Dimensionality
 - 1.1. Generate $j \times k$ matrix, \mathbf{Q} , where j indexes items, k indexes latent factors, and $q_{jk} = 1$ or 0 .
 - 1.2. First, create simple-structure \mathbf{Q}
 - 1.3. Using simple-structure \mathbf{Q} , modify to create complex-structure \mathbf{Q}^*
2. *Item Type Generation* – for each level of Model Complexity, Model Dimensionality, Test Length, and Item Type
 - 2.1. Generate $j \times 1$ vector of item difficulty values, \mathbf{B}
 - 2.2. Generate $j \times k$ matrix of item discrimination values, \mathbf{A}
3. *Latent Ability Generation* – for each level of Inter-Factor Correlation and Sample Size
 - 3.1. Generate a $k \times k$ inter-factor correlation matrix, $\mathbf{\Sigma}$, where $\Sigma_{kk} = 1.0$, Σ_{kk} is defined according to the inter-factor correlation conditions and the three correlations in the three-factor model being constrained to equality.
 - 3.2. Generate $i \times k$ matrix of latent ability distributed as multivariate normal, $\mathbf{\theta} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$
4. *Item Response Generation* – for each level of Model Complexity, Model Dimensionality, Inter-Factor Correlation, Test Length, Item Dimensionality, Item Type, and Sample Size

- 4.1. Constrain item discrimination values, \mathbf{A} , according to an element-by-element multiplication of \mathbf{Q} or \mathbf{Q}^* elements, as appropriate.
- 4.2. Generate an $i \times j$ matrix of probabilities of correct responses, \mathbf{P} , according to the 2-parameter normal-ogive multidimensional item response theory model (2-PNO MIRT; De Ayala, 2009; Lord, 1952; Reckase, 2009).
- 4.3. Generate an $i \times j$ matrix of random uniform values, \mathbf{U}
- 4.4. Generate an $i \times j$ matrix of observed dichotomous responses, \mathbf{X}
 - 4.4.1. if $P_{ij} \leq U_{ij}$ then $X_{ij} = 0$
 - 4.4.2. if $P_{ij} > U_{ij}$ then $X_{ij} = 1$

3.2. Estimation Conditions

The current study employs the weighted least squares mean- and variance-adjusted estimator (WLSMV; Muthén & Muthén, 1998-2001; Muthén, Du Toit, & Spisic, 1997) as implemented in the Mplus software package, version 6.11 (Muthén & Muthén, 1998-2010) for the estimation of models under the CFA framework. The MIRT model is also estimated using the Mplus software with similar estimation specifications as the CFA model. Mplus estimates a 2-PNO MIRT model, using the probit link function (Φ), resulting in comparable item parameters according to the transformations provided by Takane and de Leeuw (1987). Additionally, a study by Maydeu-Olivares (2001) demonstrated that parameter estimates obtained using the *Normal Ogive Harmonic Analysis Robust Method* software (NOHARM; Fraser & McDonald, 1988), which estimates the two-parameter MIRT model via an approximation to the normal ogive, were comparable to those obtained from Mplus. Default Mplus settings were typically employed, meaning that ten random sets of starting values were generated and 10

optimizations were carried out for each replication, with the exception that the number of processors was specified to take advantage of the four CPU's available on some computers used in this dissertation (PROCESSORS = 4).

3.2.1 Model Misspecification

Each of the original Q-matrices are misspecified as \mathbf{Q}' such that specific entries of \mathbf{Q} are set to $q'_{jk} = 0$ when $q_{jk} = 1$ or $q'_{jk} = 1$ when $q_{jk} = 0$. Misspecified Q-matrix entries can reflect one of three possible types: *alternate-factor* misspecification, *underfactoring*, or *overfactoring*. Alternate-factor misspecification represents instances where an item is estimated as loading on a latent factor differing from the generating factor, underfactoring represents the estimation of fewer factor loadings than specified during response generation, and overfactoring represents the estimation of more factor loadings than specified during response generation. To limit the complexity of this dissertation, alternate-factor misspecification is applied only to items demonstrating between-item multidimensionality and underfactoring is applied to items demonstrating within-item dimensionality; to limit the complexity of this dissertation as well as corresponding to previous studies of model misspecification, overfactoring is excluded from the study design.

For the *True Model* condition, no items are misspecified. For all other models, the misspecifications are pre-specified and strategically balanced within each experimental cell such that the marginal proportions of items per attribute are maintained. For models estimated according to the *Moderate Misspecification* condition, one-sixth of all items are alternate-factor misspecified; only items demonstrating between-item multidimensionality are misspecified. This means that a misspecified item is instead

estimated as an indicator of a single latent factor differing from the generating latent factor. Model estimation according to the *Severe Misspecification* condition means that one-third of all items are misspecified, which includes those that were previously misspecified under the Moderate Misspecification condition as well as an additional, equal, number of items. Under the simple-structure model condition these additional items reflect alternate-factor misspecification while misspecified items under complex-structure models reflect underfactoring. Further, correctly specified items are maintained across conditions; specific items are correctly specified regardless of the complexity or degree of misspecification of the model. The misspecified Q-matrices are presented in the Appendix.

3.2.2 Fit Indices

Model fit indices considered in this study were selected based on sensitivity to measurement model misspecification demonstrated in previous studies of measurement model misspecification, their frequency of use by practitioners, and their availability in software programs (see the review by Gierl & Mulvenon, 1995). These indices include the χ^2/df , the *Root Mean Square Error of Approximation* (RMSEA; Browne & Cudeck, 1993; Steiger & Lind, 1980), Modification Indices (MI; Sörbom, 1989), and the Wald Test (Buse, 1982). The generalized dimensionality discrepancy measure (GDDM; Levy & Svetina, 2010) and the $S\text{-}\chi^2$ (Orlando & Thissen, 2003) will be employed in the current study, representing the assessment of global- and local fit under the MIRT framework.

While the χ^2/df , RMSEA, MI, and Wald Test statistics are all commonly available in CFA and SEM software packages, the GDDM and $S\text{-}\chi^2$ were programmed by the author in the R software package (R Development Core Team, 2011). Calculation of the

GDDM is straightforward according to the formula; calculation of the $S-\chi^2$ is detailed as follows.

The $S-\chi^2$ statistic is calculated using the joint likelihood for each total score k , S_k , or the summation of all likelihoods across all distinct response patterns for each total score category. Using a recursive algorithm, the joint likelihood is computed one item at a time. Subsequently, the expected proportion of correct responses to item j under total score t , or E_{jt} , is computed as

$$E_{jt} = \frac{\iint P_j(\theta_1, \theta_2) S_{t-1}^{*j} f(\theta_1, \theta_2) \partial \theta_1 \partial \theta_2}{\iint S_t f(\theta_1, \theta_2) \partial \theta_1 \partial \theta_2}$$

where $f(\theta_1, \theta_2)$ is the population distribution of ability under a two-dimensional MIRT, S_{t-1}^{*j} is the joint likelihood for total score category $t - 1$ without item j (obtained from the recursive algorithm). The integrals in the numerator and denominator can be approximated by rectangular quadratures over the combination of equally spaced increments of θ_1 and θ_2 . The calculation of E_{jt} is generalized by expanding the integrals in the numerator and denominator to include response probability, population distributions, and derivatives with respect to k dimensions. Finally, the statistic is computed as

$$S - \chi^2 = \sum_{t=1}^T \frac{N_t (O_{jt} - E_{jt})^2}{E_{jt} (1 - E_{jt})}$$

where O_{jt} is simply the observed proportion correct for item j in total score category t and N_t is the number of examinees in total score category t .

3.2.3 Performance of Fit Statistics

Prior to analysis of the fit statistics, the model estimation process is first evaluated by examining estimation issues, commonly defined in terms of convergence failures and Heywood cases which result in negative error variances for the estimated parameters. The model estimation process is further evaluated by examining the recovery of item parameters, specifically MDIFF, MDISC, inter-factor correlations, and person estimates or θ_i , via calculations of the root mean-squared error (RMSE) and average bias. Root mean-squared error is calculated as

$$RMSE(\omega) = \sqrt{\sum_{r=1}^R (\hat{\omega}_r - \omega_r)^2 / N_R} ;$$

where ω indicates the population or generating parameter of interest, $\hat{\omega}$ is the estimated parameter, and r indexes the 250 or 1000 replications within each cell of the simulation design. This statistic describes the empirical standard error of the parameter estimates where smaller values indicate better recovery of the original, generating values. Similarly, average bias is calculated as:

$$\text{Bias}(\omega) = \frac{\sum_{r=1}^R (\hat{\omega}_r - \omega)}{N_R} ;$$

and is a signed-indicator of the magnitude of the discrepancy between the estimated and generating parameters.

With the exception of the GDDM, each of the fit indices (i.e., RMSEA, χ^2/df , MI, Wald Test, and $S\text{-}\chi^2$) is posited to follow a theoretical distribution – typically χ^2 – and, therefore, hypothesized distributional properties can be described for each, including mean, variance, skewness, kurtosis, and percentiles associated with the typical

significance levels (0.10, 0.05, 0.01). Aggregating over simulation replications, the empirical sampling distributions of the fit indices will be compared to the expected, theoretical, critical values according to the various simulation conditions. Comparison of the theoretical and empirical sampling distributions will reveal whether model complexity, model size, test length, sample size, item type, or degree of model misspecification result in violations of the assumptions of the null distribution. Suggested by authors such as Tay and Drasgow (2011), empirically-derived cut points for each fit index are then derived as the values corresponding to the 95th percentile, representing a significance level of $\alpha = 0.05$.

As stated by Fan and Sivo (2007), fit indices should be sensitive to model specification errors; sensitivity to conditions other than model misspecification is typically demonstrated as the proportion of variation in the outcome statistic attributable to the conditions resulting from a factorial analysis of variance (ANOVA). Large proportions of variance attributed to one or multiple simulation conditions indicate variability between the levels of the condition or interaction of conditions and are said to suggest sensitivity of the outcome statistic to those conditions. For each fit index a factorial ANOVA is conducted to evaluate how each model and item fit index is influenced by the various simulation conditions; the sum-of-squares attributable to a factor, or simulation condition, and the total sum of squares are used to calculate $\eta^2 = 100 \times SS_{Source}/SS_{Total}$ where η^2 represents the percentage of the sum of squares attributable to each of the experimental or simulation conditions or interactions thereof (SS_{Source}) and the total sum of squares, SS_{Total} . The current study follows a balanced design which results in orthogonal factors and the factorial ANOVA partitions the

variance of the fit indices into different components according to the simulation conditions.

In this dissertation, sensitivity is defined as $\eta^2 \geq 1.000\%$ indicating that there is a non-trivial amount of variability between the levels of the conditions. Alternately, when η^2 is smaller than 1.000% the outcome statistic is stated to be insensitive to that condition or conditions. The threshold of 1.000% has been selected for descriptive reasons, indicating a non-zero amount of variability attributable to the simulation condition. While not explicitly stated, previous research on fit index sensitivity typically discusses non-zero values of η^2 (Fan & Sivo, 2005, 2007; Jackson, 2007). The outcome statistics of interest from the successfully estimated replications are subjected to separate factorial ANOVA calculations to explore the sensitivity of each model and item fit index to test length, sample size, model complexity or item multidimensionality (depending on the unit of analysis), model size, the strength of the inter-factor correlations, and item type. Under true model specification, the sensitivity of the empirically-derived cut points will be examined as these values represent the decision points in model evaluation and should, therefore, appropriately indicate misspecified models under all simulation conditions.

The effect of model misspecification on the performance of the various item and model fit indices is of primary interest in this study. Analysis of the specific effect of degree of Model Misspecification will follow the overall procedure described earlier, considering Model Misspecification as both a factor in the ANOVA calculations as well as examining the performance of the fit indices separately according to each level of misspecification. Further, Type-I error rates and power will be evaluated for each of the fit indices. Type-I error rates for each model fit index are calculated as the proportion of

true, correctly specified, models yielding fit values falling outside the critical range; Type-I error rates for item fit indices are the proportion of correctly specified replications for which the item was judged to demonstrate poor fit. Power is assessed using the empirical cut-off values that ensure approximately nominal Type-I error rates and is computed as either the proportion of misspecified models which are correctly rejected by the model fit index or the proportion of misspecified items which are correctly rejected by the item fit index. Summaries of power for item-fit indices will be computed separately for the correctly- versus incorrectly-specified items.

3.3.Real Data Application

A real-data component is included in this dissertation to (1) to serve as an illustrative example of how the findings from the current research can be applied in practice and (2) to suggest direction and applications for future research. Item-level responses for a high-stakes grade 6 mathematics achievement assessment from a large Midwestern state were obtained via an arrangement between the state department of education and the author of this dissertation. This de-identified dataset is an early return dataset collected by the test vendor for the purpose of item calibration and early research. Additionally, this administration corresponds to test materials that have been released into the public domain by the state allowing for examination of content such as item stems and item response option.

A promotional requirement for every student in grade 6, the full achievement data set represents the population of students in the state and contained responses for 12,861 to 39 items, which include binary-scored multiple choice items, short answer items worth 2 points each, and extended response items worth 4 points each. For the purposes of this

study, the dataset is reduced to include only the 32 binary-scored items and a random sample of 1,000 examinees is drawn to represent the large sample size condition simulated in this dissertation. Since test content was available for consideration, the Q-matrix was constructed as part of an earlier research study (Gushta, Yumoto, & Williams, 2009) by assigning items to appropriate levels of the revised *Bloom's Taxonomy for Educational Objectives* (Anderson & Krathwohl, 2001; Bloom, 1956). These categories describe the cognitive processes necessary to successfully answer test items, independent of specific subject-area requirements, according to the Cognitive Process Dimension. While there are 6 categories in the Cognitive Process Dimension, only 3 were represented in this assessment: *Remembering*, which is the most basic cognitive process indicating that test items require only retrieval of stored information; *Understanding*, a more complex process requiring summarizing and comparing; and *Application*, for items requiring the use of procedures to solve familiar and novel tasks. The 2-parameter normal ogive (2-PNO) multidimensional item response theory (MIRT) model will be estimated using this data and the Q-matrix resulting from the Cognitive Process Dimensions as well as Q-matrices suggested by the content standards for this assessment as well as a Q-matrix suggested by exploratory factor analysis. The resulting values of the χ^2/df , RMSEA, GDDM, Modification Indices, $S\text{-}\chi^2$, and Wald Test fit indices are then examined according to the simulation-suggested cut points, for the purpose of selecting the most appropriate Q-matrix, identifying model or Q-matrix misspecification, and suggesting subsequent Q-matrix revision.

Chapter 4

Results of True Model Estimation

Latent variable models were estimated for dichotomous response data varying in sample size, test length, item discrimination and difficulty, difficulty (i.e., magnitude of discrepancy from average examinee ability), item multidimensionality, number of latent factors, and degree of inter-factor correlation. The current chapter presents the results of estimating correctly specified, true, models. The performance of model- and item-fit statistics estimated for these models will be used as evidence in answering the following research questions:

- 1) How similar are key percentiles (i.e., 90th, 95th, and 99th) from the empirical sampling distributions to the corresponding percentiles from the theoretical sampling distributions? In other words, how strongly do the empirical and theoretical sampling distributions differ in their upper tails?
- 2) How much do the percentiles from the empirical sampling distributions vary as a function of different test design and model conditions?
- 3) How much does the use of the percentiles from the theoretical sampling distributions inflate or deflate the nominal type-I error rate?

Additionally, the bias and precision of item and person parameters will be calculated in order to evaluate parameter recovery under true model estimation conditions. Lastly, application of the theoretical or suggested cut points are then discussed as Type I error rates.

4.1.Estimation Issues

For each of the 432 true model conditions enough replications were conducted so as to obtain 1000 successfully converged replications for each cell in the design of the simulation study. On a 64-bit dual-core 2.53GHz computer with 4.00GB of RAM the true model conditions took approximately 490 hours to complete. For the majority of the cells in the experimental design all of the 1000 replications resulted in successful estimation runs; however, 167 (38.66%) of the 432 true model conditions required additional replications with a minimum of one additional replication through to a maximum of 369 additional replications for models with 3 weakly correlated latent factors, 12 high discrimination / high difficulty items estimated as within-item multidimensional, and a sample size of 250. Table 4.1 presents the simulation conditions requiring greater than an additional 1% to achieve the necessary 1000 replications.

Table 4.1
Convergence

Test Length	Sample Size	Multi.	Item Type	2 Dimensions			3 Dimensions		
				L*	M	H	L	M	H
12	250	B	HH	101%	102%	**	125%	115%	106%
			HM				125%	111%	102%
			HL				106%	103%	
			MH				106%	102%	105%
			MM				103%	102%	102%
			ML				102%		
		W	HH	104%	103%	111%	137%	123%	113%
			HM	101%	101%	103%	119%	112%	105%
			HL				108%	104%	102%
			MH			104%	105%	103%	115%
			MM			102%	104%	102%	112%
			ML			101%			105%
24	250	B	HH			102%	103%	106%	103%
			HM					101%	
			HL					101%	
		W	HH	101%	101%	103%	106%	105%	105%
			HM				102%	102%	
			HL				102%		
			MH						101%
36	250	B	HH					102%	102%
			HM					101%	
			HL					101%	
		W	HH			102%	103%	105%	102%
			HM				102%	102%	

* Indicates inter-factor correlation: L = Low, M = Moderate, and H = High.

** 100% convergence omitted for clarity.

Generally, the proportion of replications that needed to be replaced corresponded with shorter test lengths and small sample sizes; moreover, a greater number of replications were necessary for the three-dimensional models than the two-dimensional models. These results suggest that the models are generally estimable; however, the smaller sample sizes and increased model complexity corresponded to a larger number of estimation failures and more parameters that were imprecisely estimated.

4.1.1 Results for MDIFF

Specifically, summaries of the root mean-squared error (RMSE) and average bias for MDIFF, MDISC, inter-factor correlations, and ability (i.e., θ) are presented in Table 4.2. Overall, values of the RMSE values for the MDIFF are small (mean = 0.222, median of 0.161) with the largest RMSE values corresponding to the smallest sample size ($n = 250$) but otherwise varied with respect to condition. Average bias of MDIFF is also small (mean = -0.001; median = -0.005) with the largest values occurring under the smallest sample size. Thus, overall, recovery of MDIFF parameters was mostly dependent upon sample size, though the degree of discrepancy between the true and estimated values was small across all conditions.

4.1.2 Results for MDISC

RMSE values for MDISC are slightly larger than those seen for MDIFF (mean = 0.332; median = 0.221) and the average bias values are also more positive (mean = 0.001; median = 0.003), suggesting an increased number of discrepancies of greater magnitude. The largest RMSE values are seen under conditions of the smallest sample size, shortest test length, highly discriminating items, and highly correlated latent factors.

Average bias shows behavior similar to the RMSE, though values increase as inter-factor correlation becomes stronger. Recovery of discrimination parameters is seen to be dependent on sample size, though this relationship is not straightforward.

4.1.3 Results for Inter-Factor Correlations

Inter-factor correlations across two- and three-dimensional models demonstrate small-to-moderate RMSE values, with means ranging 0.053 to 0.280 and medians of 0.048 to 0.180, where the larger values are associated with two-dimensional models. Average bias for the inter-factor correlations demonstrates similar ranges and behavior. The largest values of RMSE and average bias are associated with three-dimensional models following simple-structure, in which latent factors are highly correlated, the test length is short, and items are highly-discriminating. Further, the largest average bias values suggest that estimated inter-factor correlations are more than double the generating values.

4.1.4 Results for Person Parameter Estimation

Finally, recovery of the person parameters, alternate known as examinee ability or θ , is examined. RMSE values are small for ability across two- and three-dimensional models (mean = 0.059 to 0.072; median = 0.065 to 0.070), however, average bias is large (mean = 0.961 to 1.695; median = 0.900 to 0.965), indicating that the majority of the values were closely recovered. There are, however, many person parameter values which were poorly recovered as demonstrated in the wide range of average bias values (-19.045 to 23.364). While large RMSE values are typically associated with small sample sizes,

simple-structure three-dimensional models with highly discriminating and difficulty items, extreme average bias values follow no discernible pattern.

4.1.5 Summary of Estimation Issues

Overall, variability of parameters recovery as described by RMSE appears to be small and impacted mainly by sample size, suggesting that parameters are less precise at the smallest sample size. The magnitude of the discrepancies, indicated by average bias, is generally small for item parameters but suggests the presence of overestimated values, in the case of inter-factor correlations, and extreme values, for ability estimates, frequently associated with three-dimensional models following simple-structure with highly-discriminating items.

Table 4.2

Descriptive Statistics for Root Mean-Squared Error and Average Bias of Key Parameters

	Param.	Min	25th%	Mean	Median	75th%	Max	SD
RMSE	MDIFF	0.063	0.105	0.222	0.161	0.253	3.291	0.279
	MDISC	0.085	0.153	0.332	0.221	0.398	4.751	0.359
	ρ_{12}	0.018	0.044	0.280	0.180	0.511	0.786	0.273
	ρ_{13}	0.019	0.036	0.054	0.049	0.067	0.125	0.023
	ρ_{23}	0.019	0.035	0.053	0.048	0.066	0.121	0.023
	θ_1	0.031	0.039	0.059	0.065	0.071	0.117	0.019
	θ_2	0.032	0.050	0.072	0.070	0.086	0.194	0.030
	θ_3	0.031	0.048	0.072	0.069	0.087	0.174	0.031
Average Bias	MDIFF	-0.333	-0.031	-0.001	-0.005	0.034	0.233	0.069
	MDISC	-0.317	-0.004	0.001	0.003	0.014	0.162	0.047
	ρ_{12}	-0.354	0.000	0.508	0.409	1.010	1.821	0.528
	ρ_{13}	-0.037	0.004	0.019	0.009	0.029	0.361	0.032
	ρ_{23}	-0.616	0.039	0.209	0.218	0.365	0.904	0.274
	θ_1	-19.045	0.581	1.349	0.900	0.989	21.495	3.882
	θ_2	-11.553	0.875	1.695	0.965	0.997	23.364	3.535
	θ_3	-14.917	0.828	0.961	0.955	1.000	16.221	2.923

In a study by Finch (2011), the RMSE for item difficulty parameters estimated according to correctly specified 2-PNO models ranged 0.86 to 0.99 while RMSE for item discrimination parameters ranged 0.34 to 0.54. While the results of the current study suggest that levels of the simulation conditions affect parameter estimates and subsequent statistics dependent on these values, the parameters are generally well-recovered compared to previous research.

4.2.Distributional Characteristics of Model Fit Indices

The 90th, 95th, and 99th percentiles from the empirical sampling distributions across the 1,000 successful replications were stored and submitted to an ANOVA that included the test design and model conditions as factors. In the following, the distributional characteristics of the χ^2/df ratio, RMSEA, and GDDM model-fit indices as well as that of the $S\text{-}\chi^2$, Modification Index, and Wald Test item-fit indices are examined via descriptive statistics such as means, medians, standard deviations, and inter-quartile ranges, as well as graphically using box-and-whisker plots and empirical cumulative distribution functions for each fit index. For ease of interpretation and presentation, these statistics are summarized according to the simulation conditions for which the empirically-derived cut points, the 95th percentiles representing a nominal significance level of $\alpha = 0.05$, demonstrate sensitivity resulting from the factorial ANOVA calculations for each fit index.

The proportion of variance in the empirically-derived cut points for each model-fit index are presented as percentages in Table 4.3 according to main effects of simulation conditions and interactions thereof for which the cut points demonstrated sensitivity.

Table 4.3

Selected Percentages of Variance for Empirically-Derived Model-Fit Cut Points Under True Model Specification

Source	χ^2/df	RMSEA	GDDM
Number of Dimensions (1)	0.161	0.103	<u>6.078</u>
Test Length (2)	<u>69.925</u>	<u>40.126</u>	<u>13.885</u>
Sample Size (3)	0.193	<u>38.663</u>	<u>18.746</u>
Item Multidimensionality (4)	<u>2.125</u>	<u>1.793</u>	0.024
Inter-Factor Correlation (5)	<u>1.932</u>	<u>1.510</u>	0.006
Item Type (6)	<u>1.711</u>	0.999	<u>53.563</u>
2*3	0.816	<u>6.585</u>	<u>1.227</u>
2*6	0.404	0.510	<u>2.517</u>
3*6	<u>4.059</u>	<u>1.250</u>	<u>1.041</u>
2*3*6	<u>2.189</u>	0.583	0.078
Residuals	9.303	3.456	1.095

4.2.1 Results for χ^2/df

Large percentages of variance are attributable to test length in the empirical cut points of the χ^2/df ($\eta^2 = 69.925$) while a lesser degree of sensitivity is demonstrated to multidimensionality, inter-factor correlation, item type, the first-order interaction of test length and item type, and the second-order interaction between test length, sample size, and item type. Descriptive statistics for the χ^2/df model fit index resulting from True Model estimation are therefore presented according to test length, sample size, and item type (see the Appendix); the box-and-whisker plot shown in Figure 4.1 depicts these values graphically. As demonstrated by the medians and inter-quartile ranges, the True Models typically fit the data well resulting in values of approximately $\chi^2/\text{df} = 1$. The interaction of test length and sample size is clear, however, in the ranges of χ^2/df and the corresponding fluctuations in the 90th, 95th, and 99th percentiles. The empirical cumulative distribution functions for the χ^2/df are depicted in Figure 4.2 according to the

same conditions where they are seen to deviate from the theoretical distribution⁴. The distribution of this fit statistic most closely approaches the theoretical distribution under the short test length condition and shows increasing deviation from the theoretical distribution as the test length increases, more closely approximating 1.000. Overall, positive skewness in the distribution of the χ^2/df indicates that the suggested cut points of 2 or 3 (Byrne, 1989; Carmines & McIver, 1981; Marsh & Hocevar, 1985) are inappropriate for the conditions presented in this study as they are much larger than the empirically-derived cut points corresponding to the 90th, 95th, and 99th percentiles.

⁴ Values for the theoretical distribution of χ^2/df can be fully determined by sample size and test length.

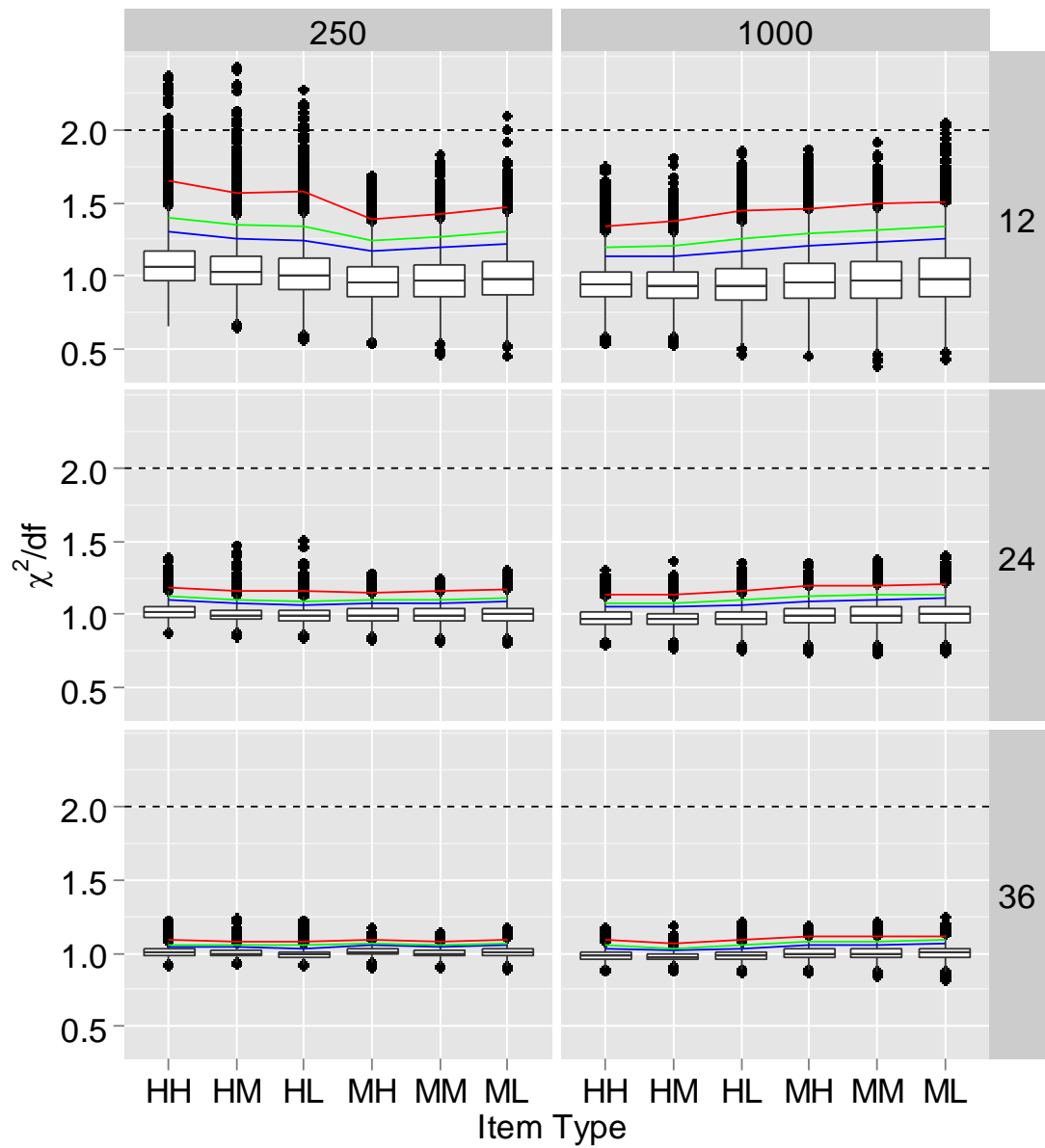


Figure 4.1. Box-and-whisker plot for the χ^2/df ratio.

Presented according to item type, test length (rows), and sample size (columns). The solid lines represent the 90th percentile (blue), 95th percentile (green), and 99th percentile (red); the dashed line represents the most conservative suggested cut point ($\chi^2/\text{df} = 2$).

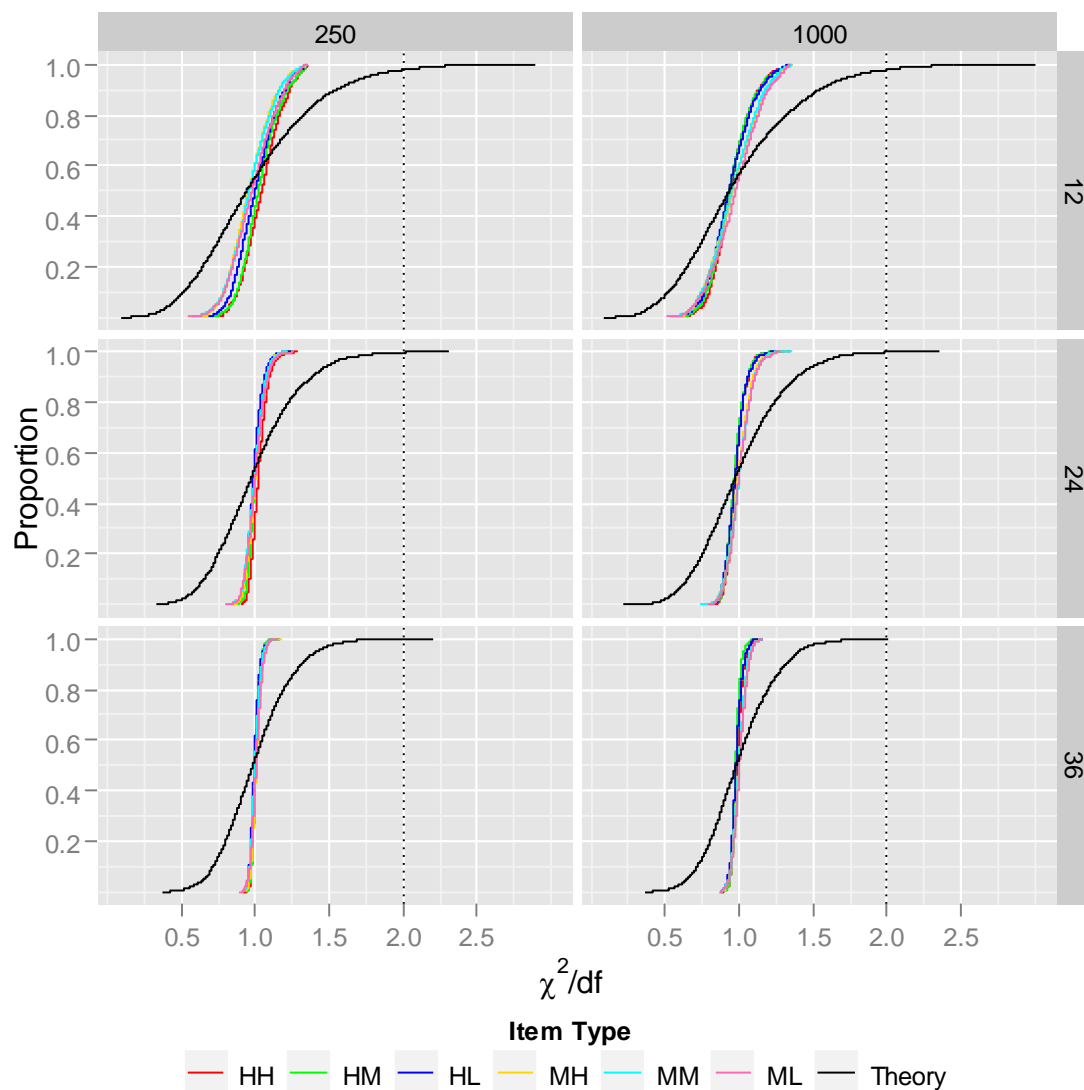


Figure 4.2. Cumulative distribution functions for the χ^2/df ratio.

Presented according to item type, test length (rows), and sample size (columns). The black line represents the theoretical distribution; the dashed line represents the most conservative suggested cut point ($\chi^2/\text{df} = 2$).

4.2.2 Results for RMSEA

Most importantly, values for the RMSEA index under correct model specification are frequently close to 0 as one would theoretically expect. Similar to the χ^2/df , the largest percentage of variance in the RMSEA empirical cut points is attributable to test length ($\eta^2 = 40.126$) with substantial variance also attributable to sample size ($\eta^2 = 38.663$) and the interaction of these two simulation conditions ($\eta^2 = 6.585$). The RMSEA demonstrates very little sensitivity to the conditions of multidimensionality, inter-factor correlation, and the interaction of sample size and item type. Descriptive statistics for the RMSEA model fit index under True Model estimation are, therefore, presented according to test length, sample size, and multidimensionality in the Appendix and as box-and-whisker plots in Figure 4.3. The RMSEA values reflect that the True Models fit the data well, as the median and inter-quartile ranges approximate 0.000 and values corresponding to the 90th, 95th, and 99th percentiles range from 0.007 to 0.049. Values of the RMSEA typically decrease with sample size and test length; decrease due to sample size is pronounced though less noticeable as test length decreases. The modest effect of multidimensionality can be seen under the simple-structure as RMSEA values demonstrate greater variability than under complex-structure. Overall, approximately half of all replications resulted in RMSEA values approaching zero. It must be noted that RMSEA values of zero do not necessarily indicate perfect fit but only a degree of misfit smaller than the precision of the software is able to detect.

The empirical cumulative distribution functions (ECDFs) for the RMSEA are depicted against the theoretical distribution in Figure 4.4 separately for different test length, sample size, and dimensionality conditions. Severe positive skewness is

demonstrated in these graphs which suggest that the empirical distributions of the RMSEA do not follow the theoretical distribution and are strongly influenced by the large proportion of RMSEA values estimated to be zero; Therefore, comparing the ECDFs against the suggested cut points of $RMSEA = 0.05$ or 0.06 (Hu & Bentler, 1999) indicates that the empirical cut points differ under a number of conditions and are not well represented or approximated by the suggested, static, cut points which are typically much larger than the 90th, 95th, and 99th percentiles.

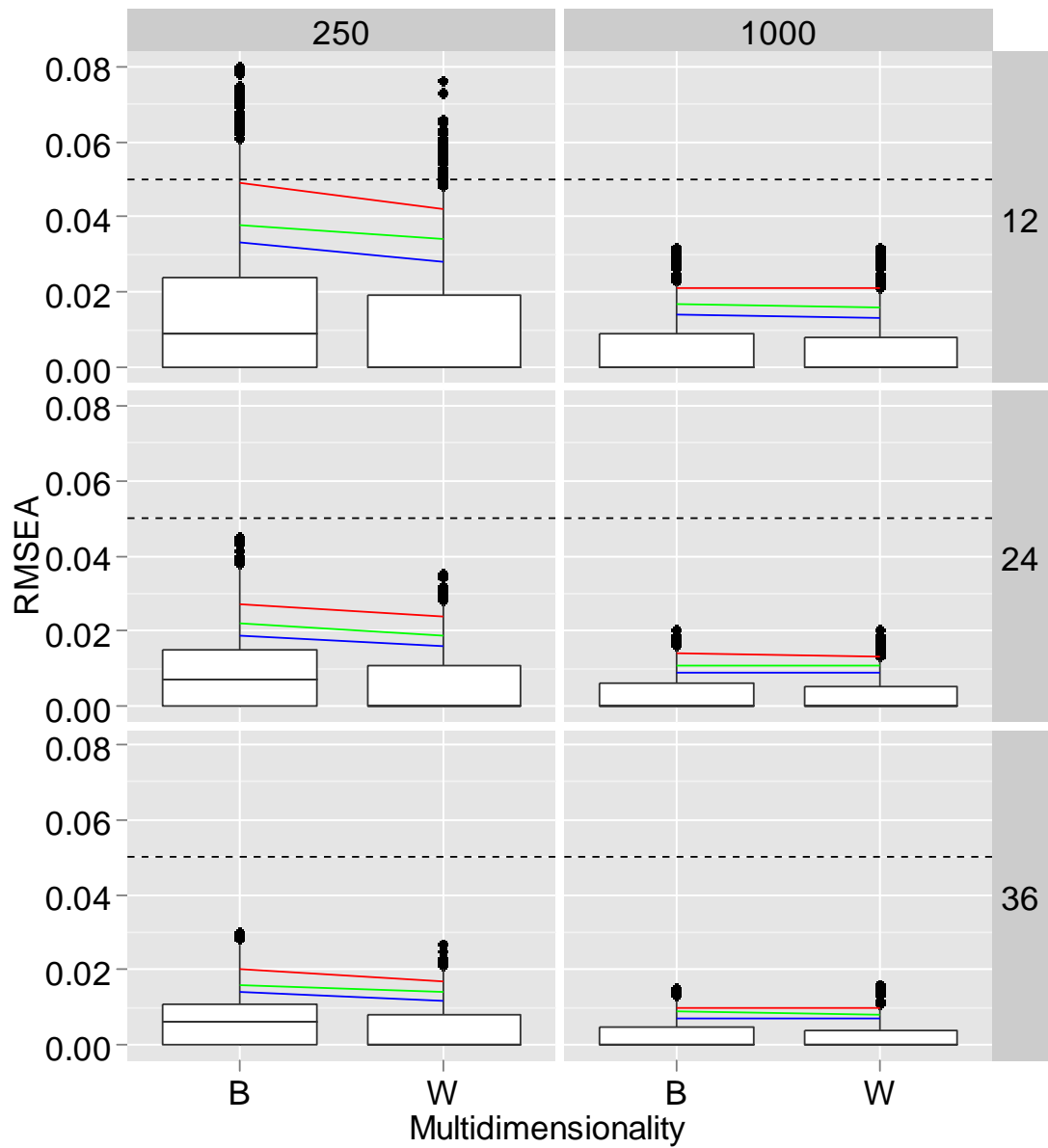


Figure 4.3. Box-and-whisker plots for RMSEA.

Presented according to multidimensionality, test length (rows), and sample size (columns). The solid lines represent the 90th percentile (blue), 95th percentile (green), and 99th percentile (red); the dashed line represents the most conservative suggested cut point (RMSEA = 0.05).

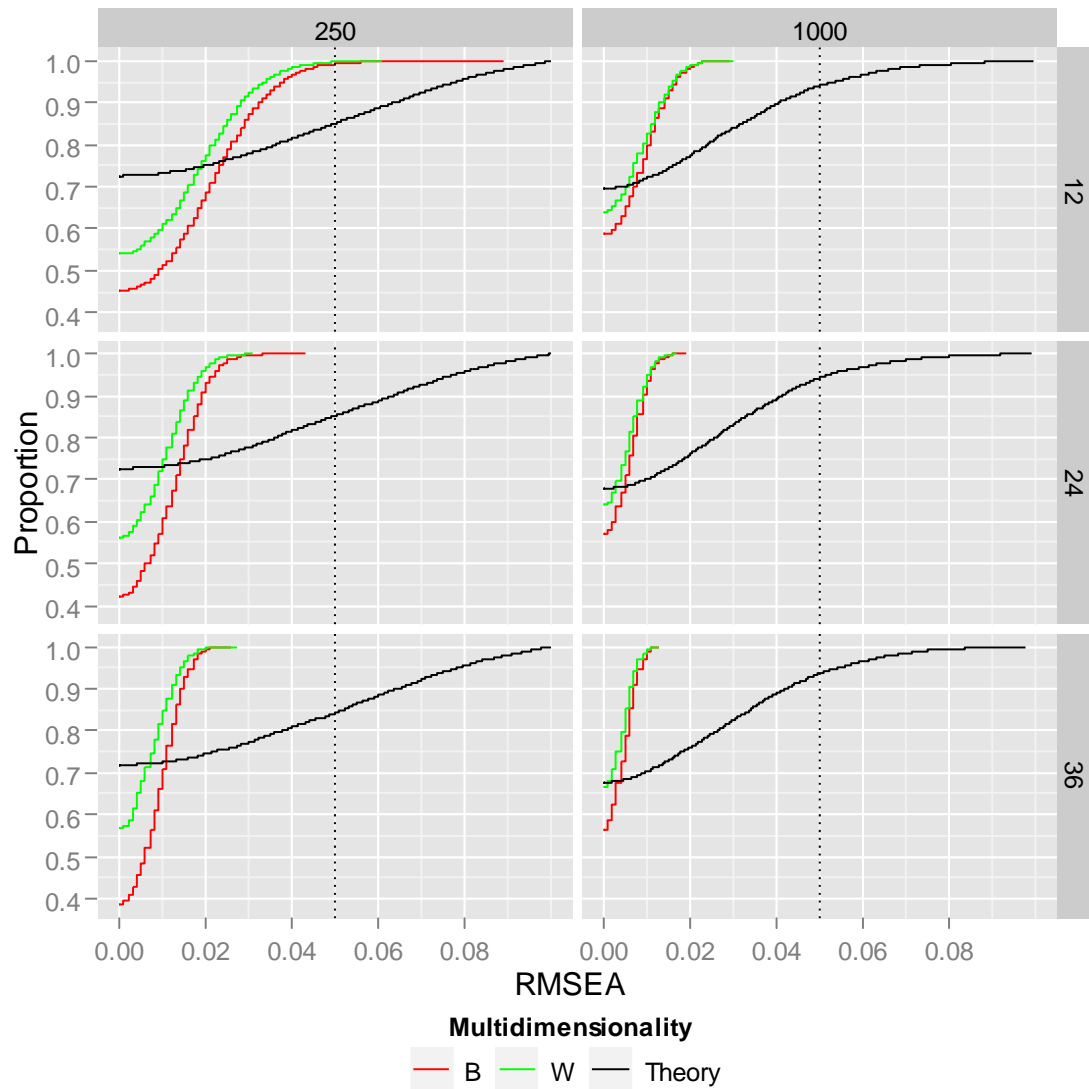


Figure 4.4. Empirical cumulative distribution functions for the RMSEA.

Presented according to multidimensionality, test length (rows), and sample size (columns). The theoretical distribution is displayed as a black line; the most conservative suggested cut point is displayed as a dashed line (RMSEA = 0.05).

4.2.3 Results for GDDM

Most importantly, values for the GDDM under correct model specification are numerically very close to 0 as one would theoretically expect. Nevertheless, a follow-up analysis of the variation of the values for the GDDM was conducted to further describe the trends in these values. Unlike the other model-fit indices, empirical cut points for the GDDM demonstrate the greatest sensitivity to item type ($\eta^2 = 16.133$) while also being sensitive to sample size ($\eta^2 = 18.746$) and test length ($\eta^2 = 13.885$). The last of the descriptive statistics for model-fit indices are also presented in the Appendix, according to item type, sample size, and test length. From these statistics and the box-and-whiskers plots illustrating the descriptive statistics (Figure 4.5), it is seen that values of the GDDM decrease substantially with both item discrimination and item difficulty. Additionally, GDDM values for the empirical cut points decrease with test length, especially when sample sizes are large; under small sample sizes, values of the GDDM reduce less drastically by test length. GDDM cut points are smallest when 36 high-discrimination / high-difficulty items are estimated using a sample size of $n = 1000$.

While the GDDM follows no known theoretical distribution, the empirical cumulative distribution functions are plotted in Figure 4.6 according to test length, sample size, and item type. The effect of item type and test length is evident as values of the GDDM decrease with item discrimination, difficulty, and test length, as seen in the box-and-whisker plots. There are no suggested or theoretical cut points against which to compare the 90th, 95th, and 99th percentiles.

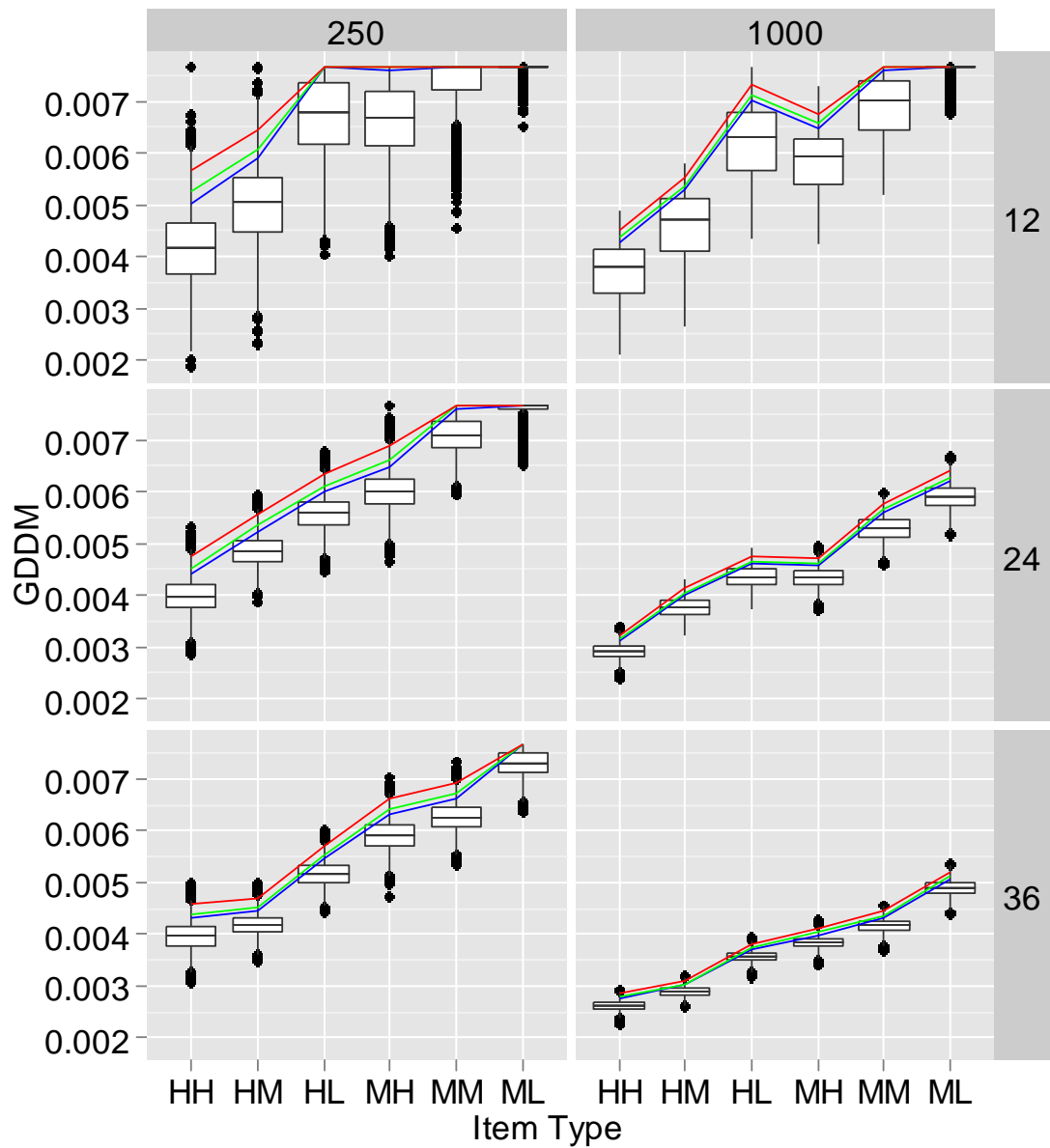
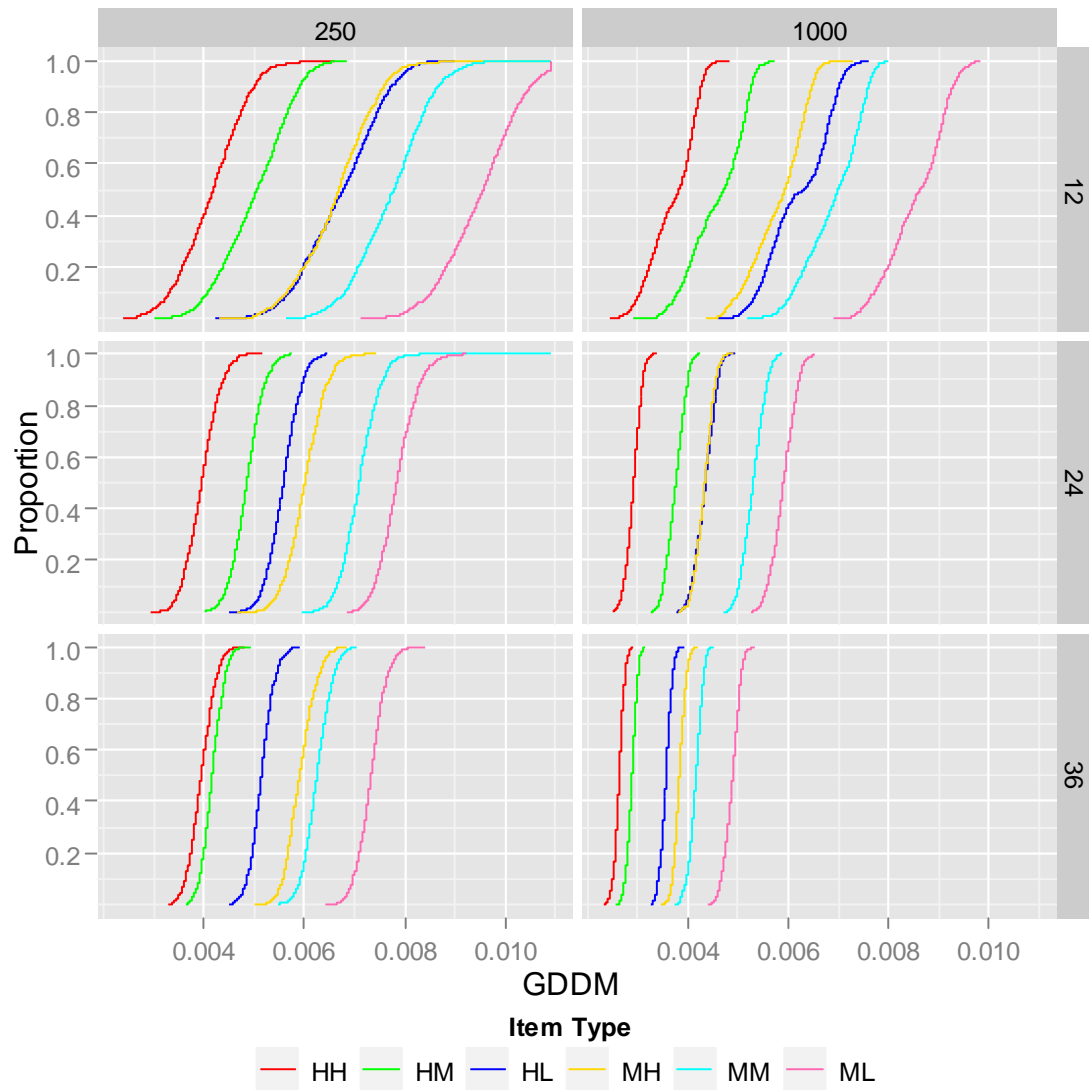


Figure 4.5. Box-and-whisker plots for GDDM.

Presented according to item type, test length (rows), and sample size (columns). The solid lines represent the 90th percentile (blue), 95th percentile (green), and 99th percentile (red). Outliers excluded for clarity.



*Figure 4.6. Empirical cumulative distribution functions for the GDDM.
Presented according to item type, test length (rows), and sample size (columns).*

4.3. Distributional Characteristics of Item Fit Indices

Main effects and specific interactions for the $S-\chi^2$, Modification Index, and Wald Test item-fit indices, in which the empirically-derived cut points demonstrated sensitivity resulting from the factorial ANOVAs are presented in Table 4.4 as percentages.

Table 4.4

Selected Percentages of Variance for Item-Fit Statistics by Simulation Condition Under True Model Specification

Source	$S-\chi^2$	Modification Index			Wald Test		
		1	2	3	1	2	3
Dimensions (1)	0.039	10.954	10.252		0.931	0.004	
Test Length (2)	35.624	9.325	9.914	9.921	2.600	2.081	3.432
Sample Size (3)	21.280	16.899	15.888	22.029	15.694	12.417	18.597
Item Multidm. (4)	10.177	0.448	0.532	0.954	54.543	61.438	48.333
Inter-factor Corr. (5)	11.199	37.758	38.188	40.368	0.423	0.345	0.514
Item Type (6)	0.411	9.246	9.146	10.827	10.750	10.208	14.303
1*2	0.108	1.221	0.567		0.012	0.104	
2*3	1.722	4.196	4.358	4.627	0.173	0.107	0.183
2*5	0.779	3.254	3.150	3.518	0.071	0.054	0.150
2*6	0.625	1.076	1.446	1.128	0.204	0.113	0.222
3*4	1.478	0.026	0.001	0.254	3.836	3.349	2.752
3*5	1.850	0.309	0.607	0.515	0.044	0.028	0.056
3*6	1.187	0.534	0.848	0.868	0.749	0.699	1.222
4*6	2.418	0.070	0.135	0.159	4.660	5.948	5.139
5*6	2.429	0.553	0.688	1.447	0.015	0.015	0.025
3*4*5	1.349	0.004	0.007	0.087	0.117	0.071	0.249
Residuals	1.560	1.542	1.378	1.220	0.362	0.449	0.578

4.3.1 Results for $S-\chi^2$

The 95th percentiles of the $S-\chi^2$ item fit index resulting from True Model estimation demonstrate sensitivity to a great number of main effects and interactions, largest among these are the sensitivity to test length ($\eta^2 = 35.624$), sample size ($\eta^2 = 21.280$), and inter-factor correlation ($\eta^2 = 11.199$). Descriptive statistics for the $S-\chi^2$ item-fit index are provided in the Appendix according to these conditions and graphically

depicted as box-and-whisker plots in Figure 4.7. Values of the $S-\chi^2$ range from approximately zero to 35 across conditions; the range and magnitude of $S-\chi^2$ values increases with sample size and test length. Further, this item-fit index shows an effect of inter-factor correlation under large sample sizes as values of $S-\chi^2$ increase with the degree of inter-factor correlation.

While the $S-\chi^2$ appears to roughly approximate the theoretical χ^2 distribution (Figure 4.8), under small sample sizes for the longest test, the empirical cumulative distribution function increasingly deviates from the theoretical distribution under the larger sample size and smaller test lengths. Deviation from the theoretical distribution is also induced by strong inter-factor correlation.

These aggregate descriptive statistics cannot be compared to theoretical cut points as the degrees of freedom for the $S-\chi^2$ are specific to each item and set of simulation conditions based on the number of valid observed score categories. However, noting that the cut point for one degree of freedom is $\chi^2 = 3.841$ and the cut point for 35 degrees of freedom, the maximum possible, is $S-\chi^2 = 49.801$, the theoretical cut points always exceed the empirical values for small sample sizes while they may approximate the 95th percentile under large sample sizes when latent factors are highly correlated.

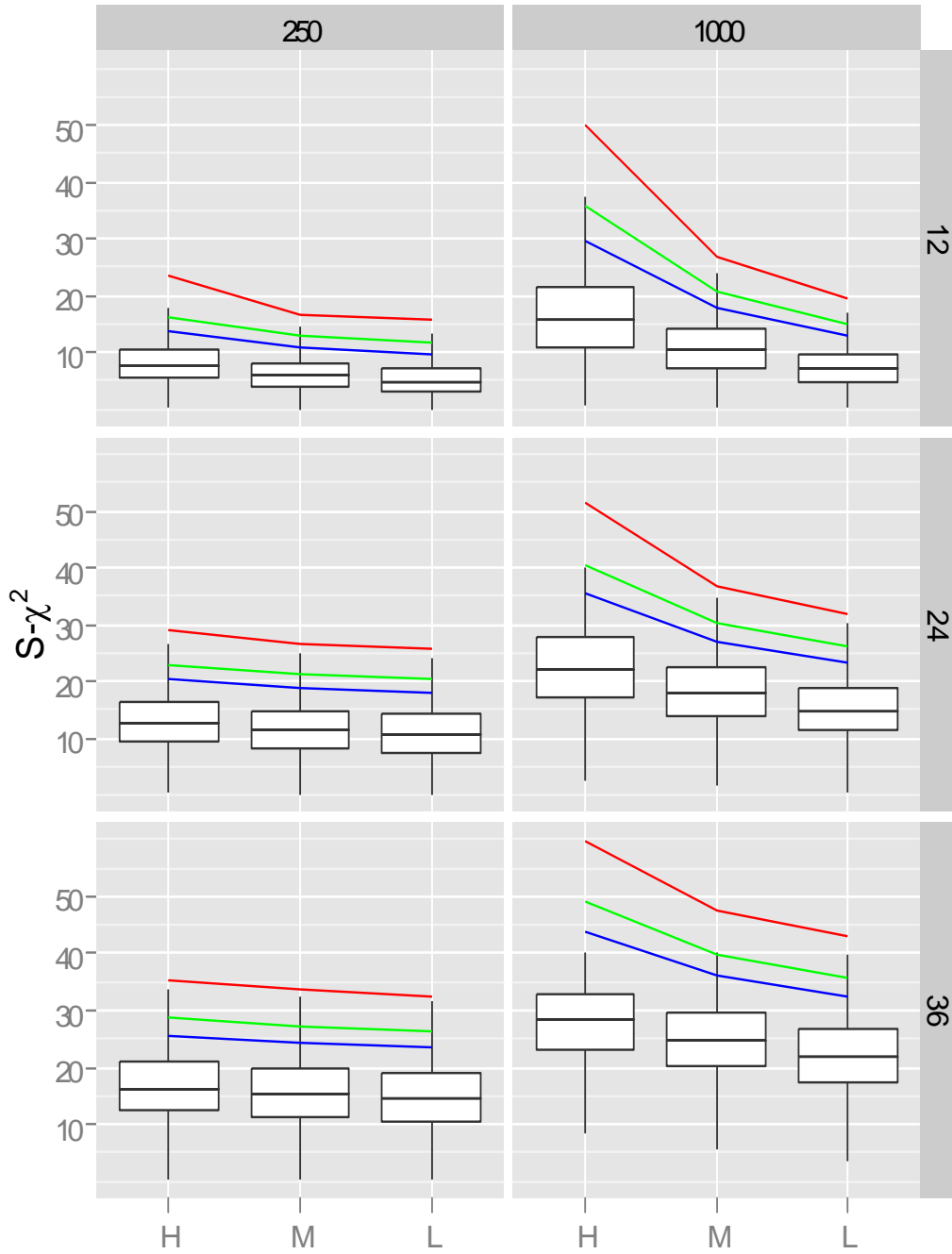


Figure 4.7. Box-and-whisker plots for $S\text{-}\chi^2$.

Presented according to inter-factor correlation, test length (rows), and sample size (columns). The solid lines represent the 90th percentile (blue), 95th percentile (green), and 99th percentile (red). Outliers have been omitted for clarity.

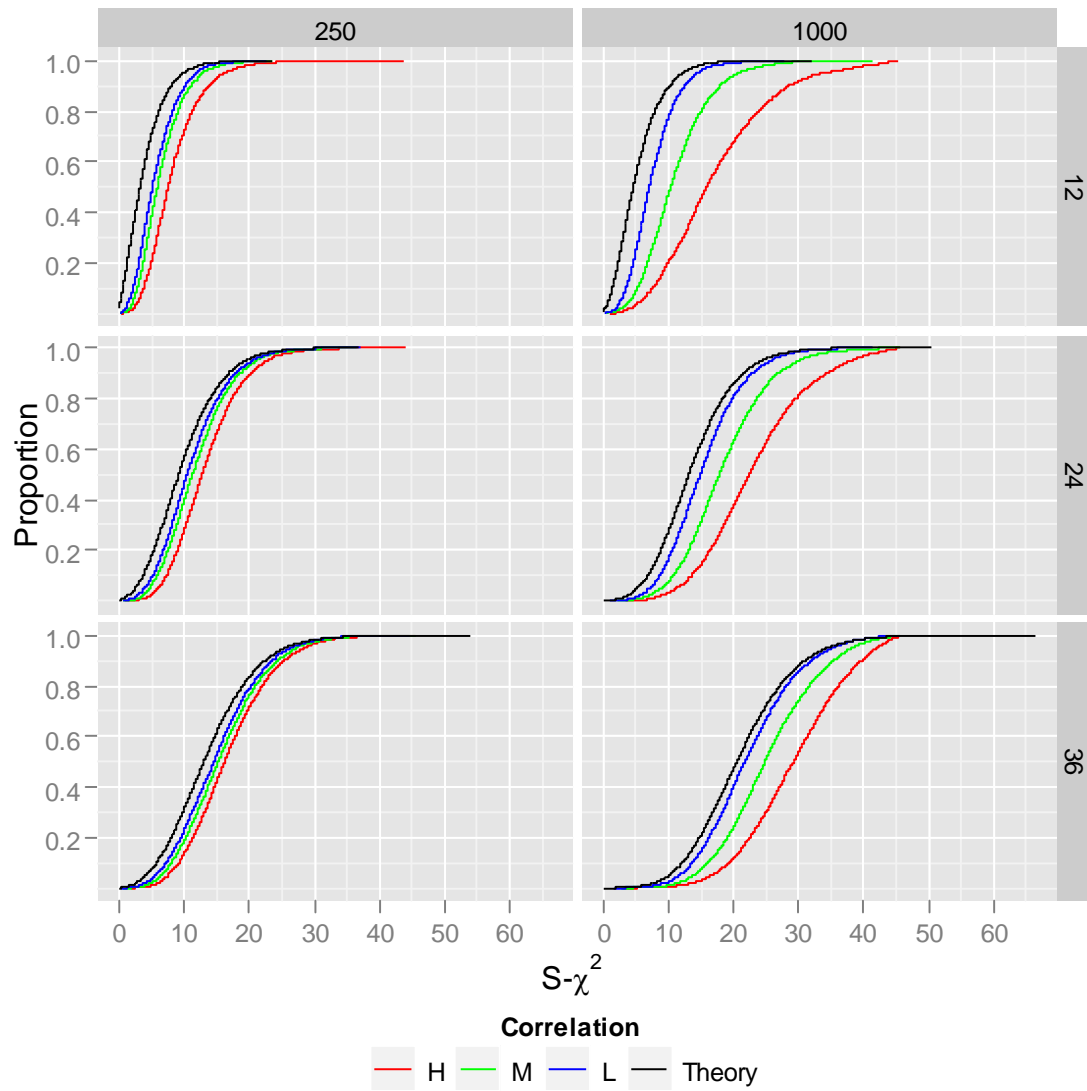


Figure 4.8. Empirical cumulative distribution functions for the $S-\chi^2$.

Presented according to item type, test length (rows), and sample size (columns). The theoretical distribution is displayed as a black line.

4.3.2 Results for Modification Indices

The 95th percentiles of the Modification indices across all three latent factors demonstrate greatest sensitivity to inter-factor correlation ($\eta^2 = 37.758$ to 40.368) and sample size ($\eta^2 = 15.88$ to 22.029). Modification indices for latent factors 1 and 2 next demonstrate sensitivity to test length ($\eta^2 = 9.325$ to 9.914) while Modification Index 3 is next-most sensitive to item type ($\eta^2 = 10.827$), though the magnitude of difference from the test length factor ($\eta^2 = 9.921$) is very small, a difference that is likely the result of removing number of dimensions from the ANOVA since Modification Index 3 can only be estimated for models with three latent factors.

Descriptive statistics for all three Modification indices are presented in the Appendix according to inter-factor correlation, sample size, and test length. The box-and-whisker plots for these three item-fit indices are presented in Figure 4.9. Considering the descriptive statistics together with the sensitivity analysis results, it is apparent that the Modification indices perform similarly regardless of the dimension for which the statistic was estimated; subsequently, only Modification Index 1 will be discussed as representative of all three values.

Values of the Modification Index approximate zero and are typically less than 5.000, indicating that items are estimated as loading correctly on the associated latent factor, though the range of values decreases as inter-factor correlation increases. Additionally, values demonstrate an increase in magnitude and variability with larger sample sizes while decreasing with additional latent factors. At a nominal significance of 0.05, the theoretical cut point for the Modification Index is $\chi^2 = 3.841$ with one degree of freedom which overestimates some of the empirical percentiles representing usual

nominal significance values but more often underestimates values of the empirically-derived cut scores, indicating that a greater proportion of items would be identified as misspecified as a result of using the theoretical cut points.

Empirical cumulative distribution functions for all three Modification indices are presented in Figure 4.10 according to inter-factor correlation, number of dimensions, and sample size. Generally, values of this fit index appear to well-approximate the theoretical distribution when the inter-factor correlation is high. As the degree of correlation decreases, however, the empirical distributions demonstrate increasing negative skewness and deviation from the theoretical distribution. This deviation is amplified under the larger sample size condition and models with three latent factors.

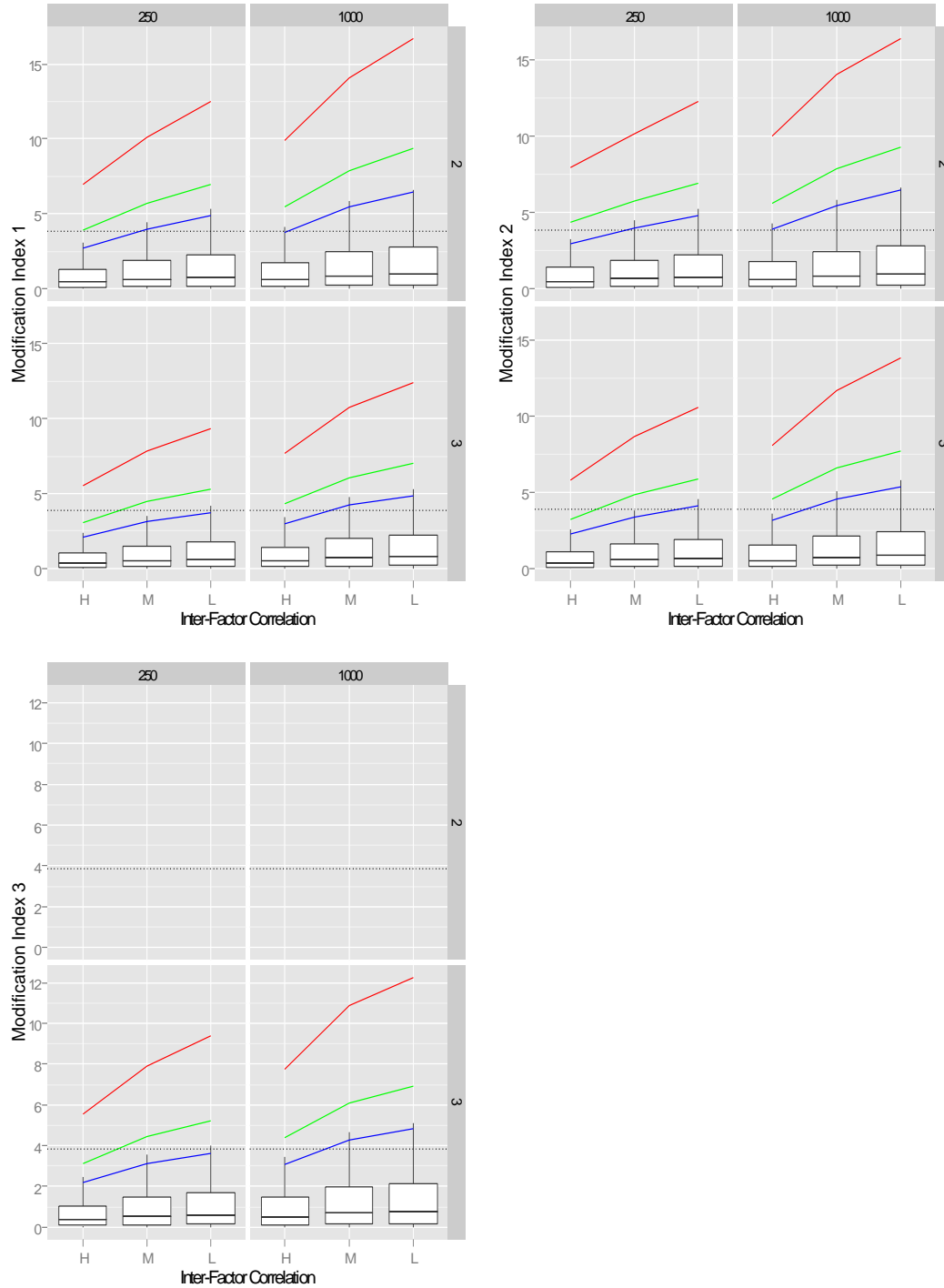


Figure 4.9. Box-and-whisker plots for the Modification Indices.

Presented according to inter-factor correlation, number of dimensions (rows), and sample size (columns). The solid lines represent the 90th percentile (blue), 95th percentile (green), and 99th percentile (red); the dotted line indicates the theoretical cut point. Outliers have been omitted for clarity.

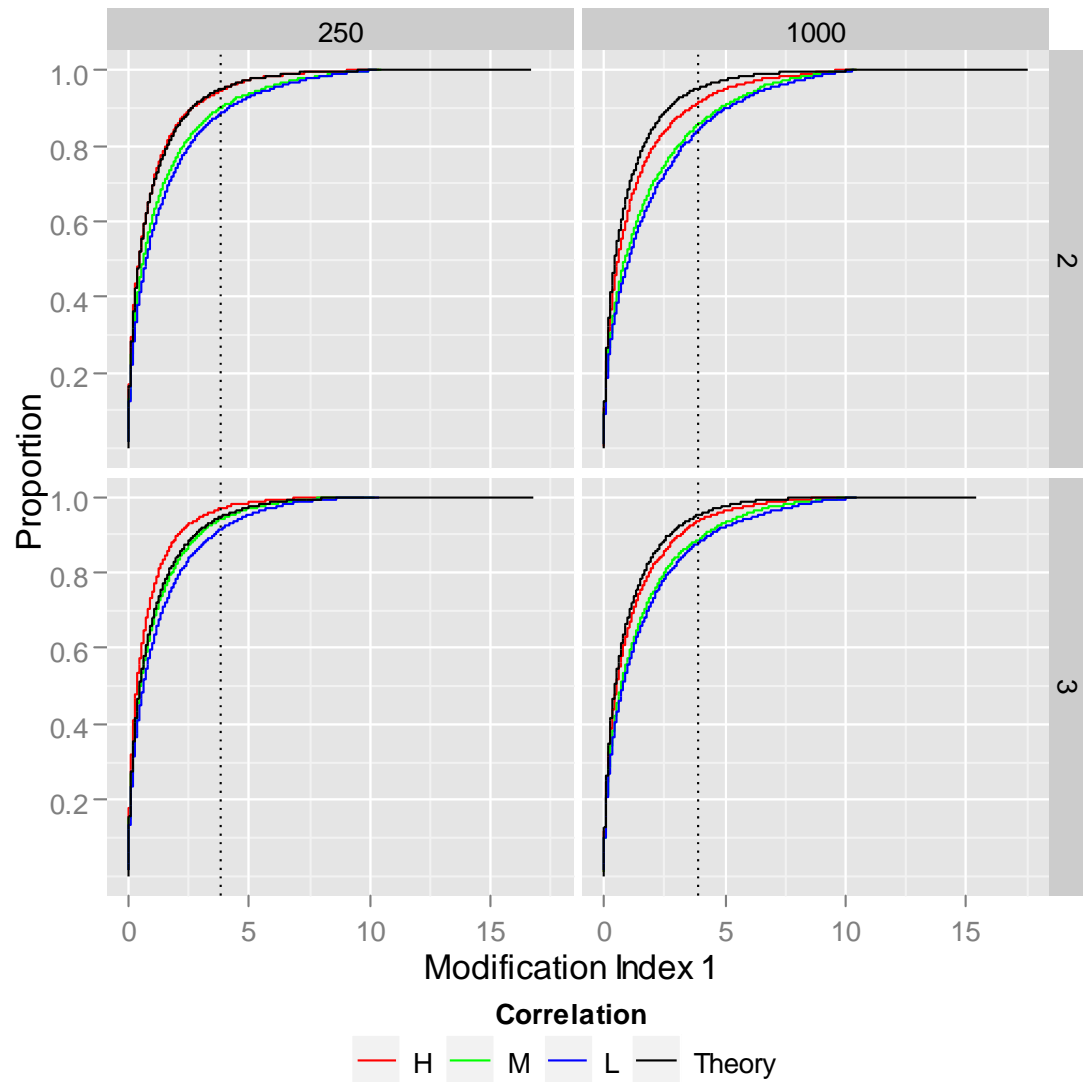


Figure 4.10. Empirical cumulative distribution functions for the Modification Index on latent factor 1.

By inter-factor correlation, number of dimensions (rows), and sample size (columns). The theoretical distribution is displayed as a black line; the theoretical cut point ($MI = 3.841$) is displayed as a dashed line.

4.3.3 Results for Wald Test Statistics

Wald Test statistics indicate significance of a freely estimated factor loading in a confirmatory factor model; therefore, there is a Wald Test value for each factor that an item is associated with; between-item multidimensionality results in a single Wald Test value while within-item multidimensionality as defined in this study results in two Wald Test values. As an indicator of significance for the estimated factor loading, critical values for the Wald Test indicate the lower bound necessary for a parameter to be considered as correctly estimated. Unlike the other fit indices, Wald Test values smaller than the critical values indicate misspecification; therefore, empirically-derived cut points are calculated for the 10th, 5th, and 1st percentiles.

The patterns of sensitivity in the 95th percentiles of the Wald Test values are similar across the three dimensions, therefore, discussion will refer to the Wald Test in general rather than the values associated with a particular latent factor. The Wald Test demonstrates greatest sensitivity to item multidimensionality (η^2 ranges 48.333 to 61.438), sample size (η^2 ranges 12.417 to 18.597), and item type (η^2 ranges 10.208 to 14.303). Descriptive statistics for all three Wald Test indices are presented in the Appendix according to these simulation conditions and the box-and-whisker plots are presented in Figure 4.11. Values of the Wald Test range approximately zero to 50 for between-item multidimensionality, increasing with sample size and discrimination while decreasing with difficulty. Values of the Wald Test under within-item multidimensionality range approximately zero to 20 and demonstrate similar patterns as under within-item dimensionality though less extreme and within a more restricted range. For sample sizes of 1000 and within-item multidimensionality the theoretical cut point,

calculated as a $\chi^2 = 3.841$ with one degree of freedom, approximates the empirically-derived cut point (i.e., 5th percentile), however, the theoretical cut point approximates the median Wald Test value under the smaller sample size while largely underestimating the distribution of values when items are between-item multidimensional.

Empirical cumulative distribution functions for the Wald Test (on latent factor 1) are presented according to item type, item multidimensionality, and sample size in Figure 4.12. It appears that the observed values of the Wald Test do not follow the theoretical χ^2 distribution with one degree of freedom; except when sample sizes are small and items are estimated as within-item multidimensional, otherwise values of the Wald Test are typically much larger than expected. Large sample sizes and tests comprised of highly-discriminating between-item multidimensional items show the greatest deviation of Wald Test values from the theoretical distribution.

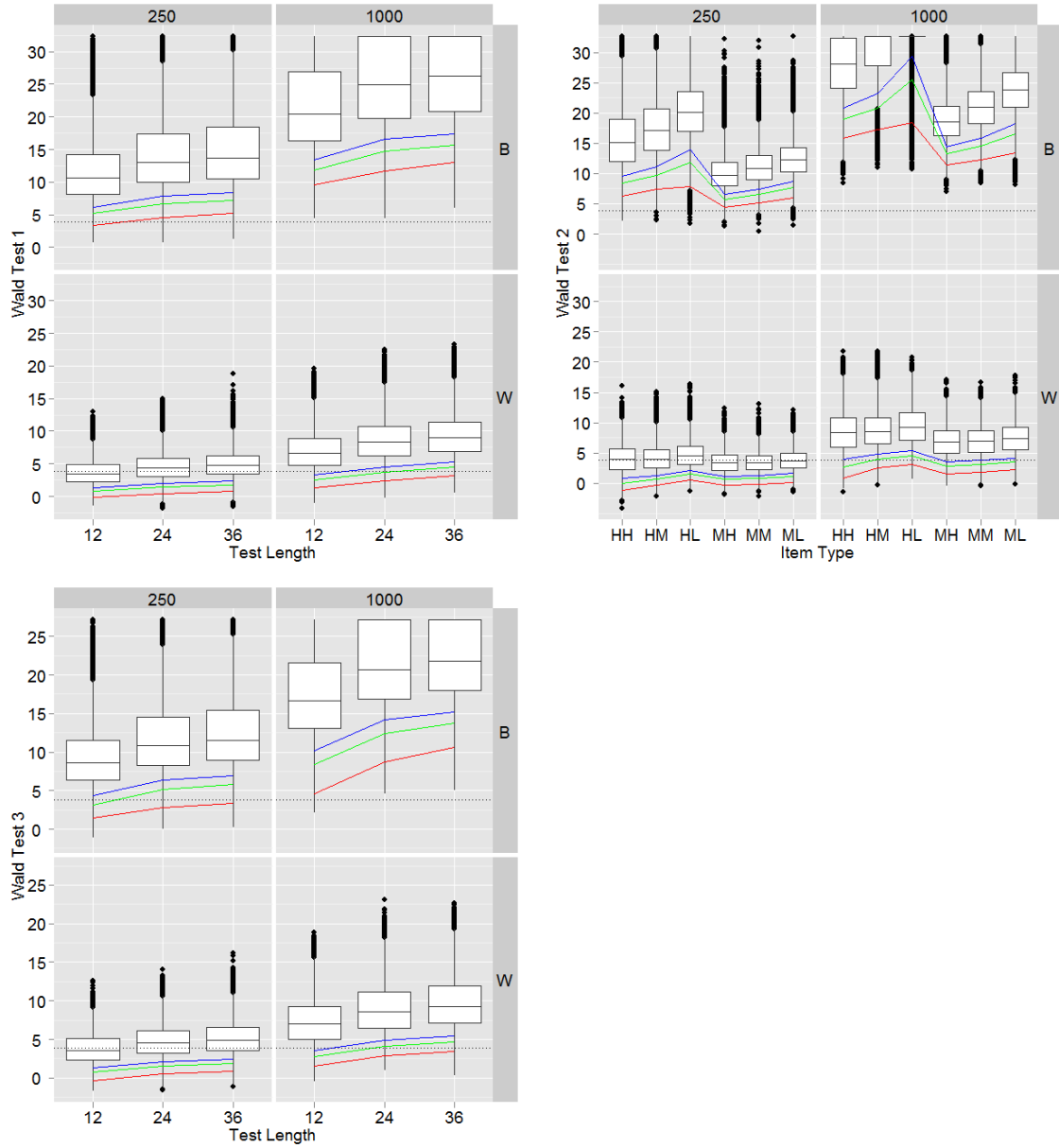


Figure 4.11. Box-and-whisker plots for the Wald Tests.

Presented according to item type or test length, multidimensionality (rows), and sample size (columns). The solid lines represent the 10th percentile (blue), 5th percentile (green), and 1st percentile (red); the dotted line indicates the theoretical cut point.

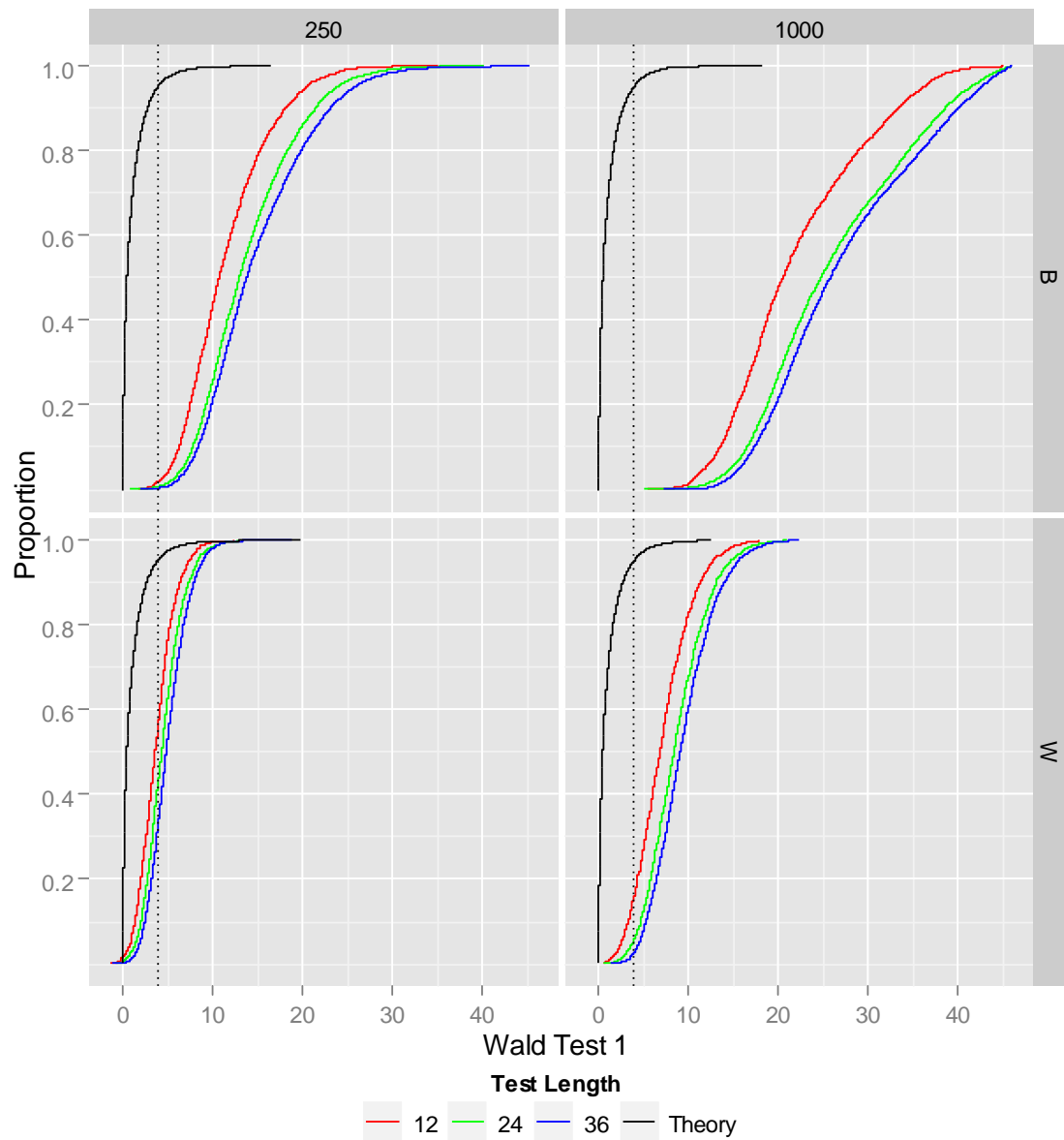


Figure 4.12. Empirical cumulative distribution functions for the Wald Test on latent factor 1

Presented according to test length, multidimensionality (rows), and sample size (columns). The theoretical distribution is displayed as a dotted line; the theoretical cut point is a dashed line.

4.4. Estimation Bias for Type-I Error Rates under Theoretical Sampling Distributions

Based on the previously observed discrepancies between theoretical and empirical sampling distributions, if theoretical cut points (e.g., χ^2 with one degree of freedom for the Modification Index and Wald Test) were employed in evaluating correctly specified models, the actual type-I error rate would differ from the nominal type-I error rate. Similarly, it is interesting to explore from a hypothesis-testing perspective what type-I error rates for the RMSEA would be like if cut points suggested by previous research (e.g., Byrne, 1989; Carmines & McIver, 1981; Hu & Bentler, 1999; Marsh & Hocevar, 1985) were incorrectly perceived as being associated with hypothesis testing, rather than effect size quantification.

Actual type-I error rates resulting from the application of the most conservative suggested cut point to the χ^2/df ratio model fit index ($\chi^2/\text{df} = 2.0$) are presented as box-and-whisker plots in Figure 4.13 according to item type, test length, and sample size – the same conditions to which empirical sensitivity was demonstrated. The suggested cut point results in underestimation of the Type-I error rate for all simulation conditions. Generally, the suggested cut point fails to reject any of the models, evidenced by median values approximating zero, with Type-I error rates approaching 0.01 under small sample sizes of short tests comprised of highly-discriminating items.

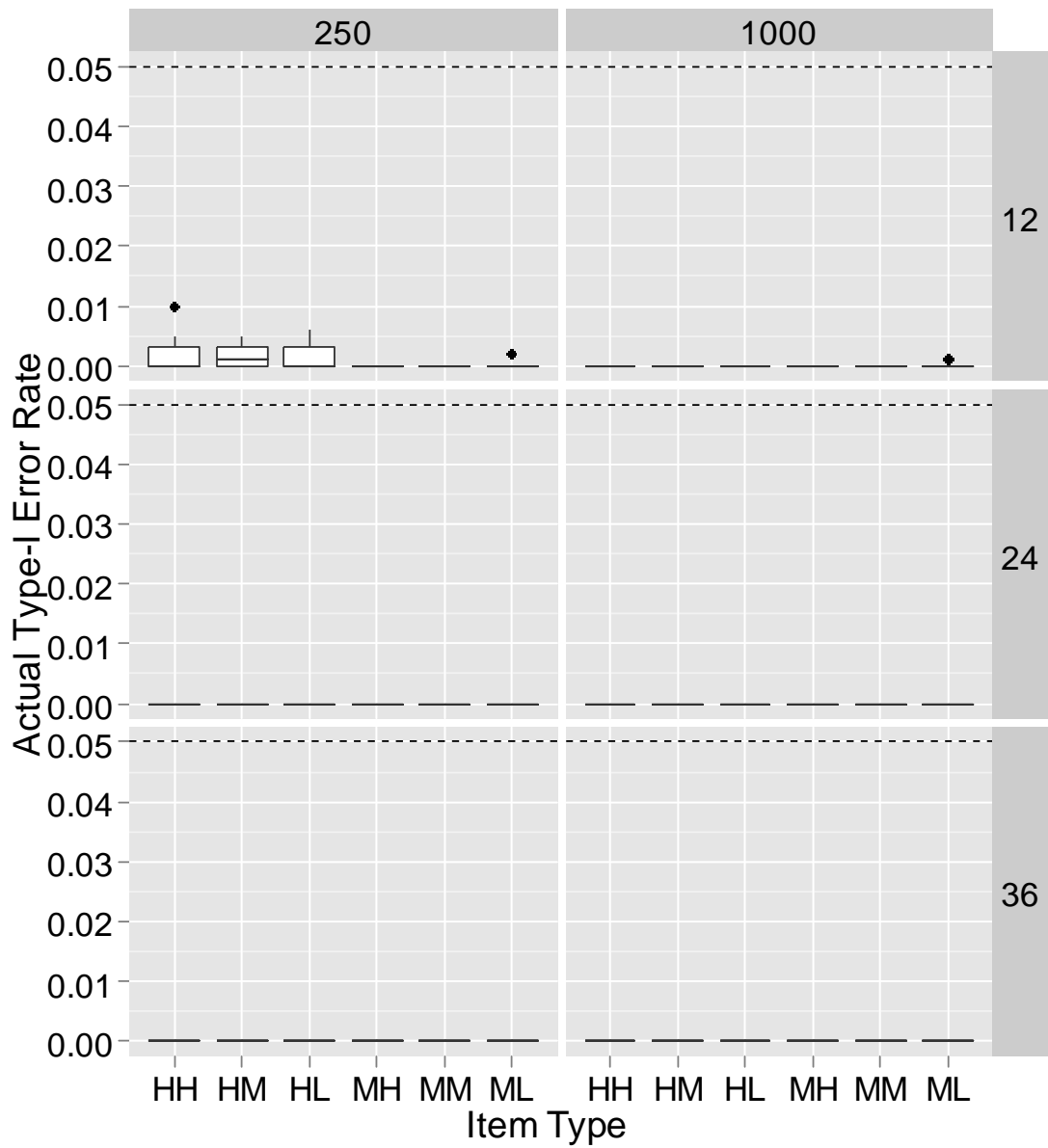


Figure 4.13. Actual Type-I error rates for the χ^2/df ratio.

Results are evaluated against the most conservative theoretical cut point ($\chi^2/df = 2.0$). Displayed according to item type, test length (rows), and sample size (columns). The dotted line indicates the nominal significance level

Figure 4.14 shows similar results when models are evaluated against the suggested RMSEA cut point of 0.05, though small samples and short tests following simple-structure show increased Type-I error rates – approaching the expected nominal level of 0.05. In other words, correct models often have RMSEA values much lower than 0.05, which means that they would certainly be considered as well-fitting which is desirable from a descriptive perspective. From a hypothesis-testing perspective this technically does not ensure nominal type-I error rates, however, for which a finer differentiation of RMSEA values closer to 0 under different test design conditions is necessary. This situation is very similar for the GDDM, which is a discrepancy measure where a GDDM of 0 indicates perfect model-data fit. While values close to 0 are desirable, a finer differentiation of values closer to 0 under different test design conditions is necessary if the GDDM is to be used within a hypothesis-testing framework.

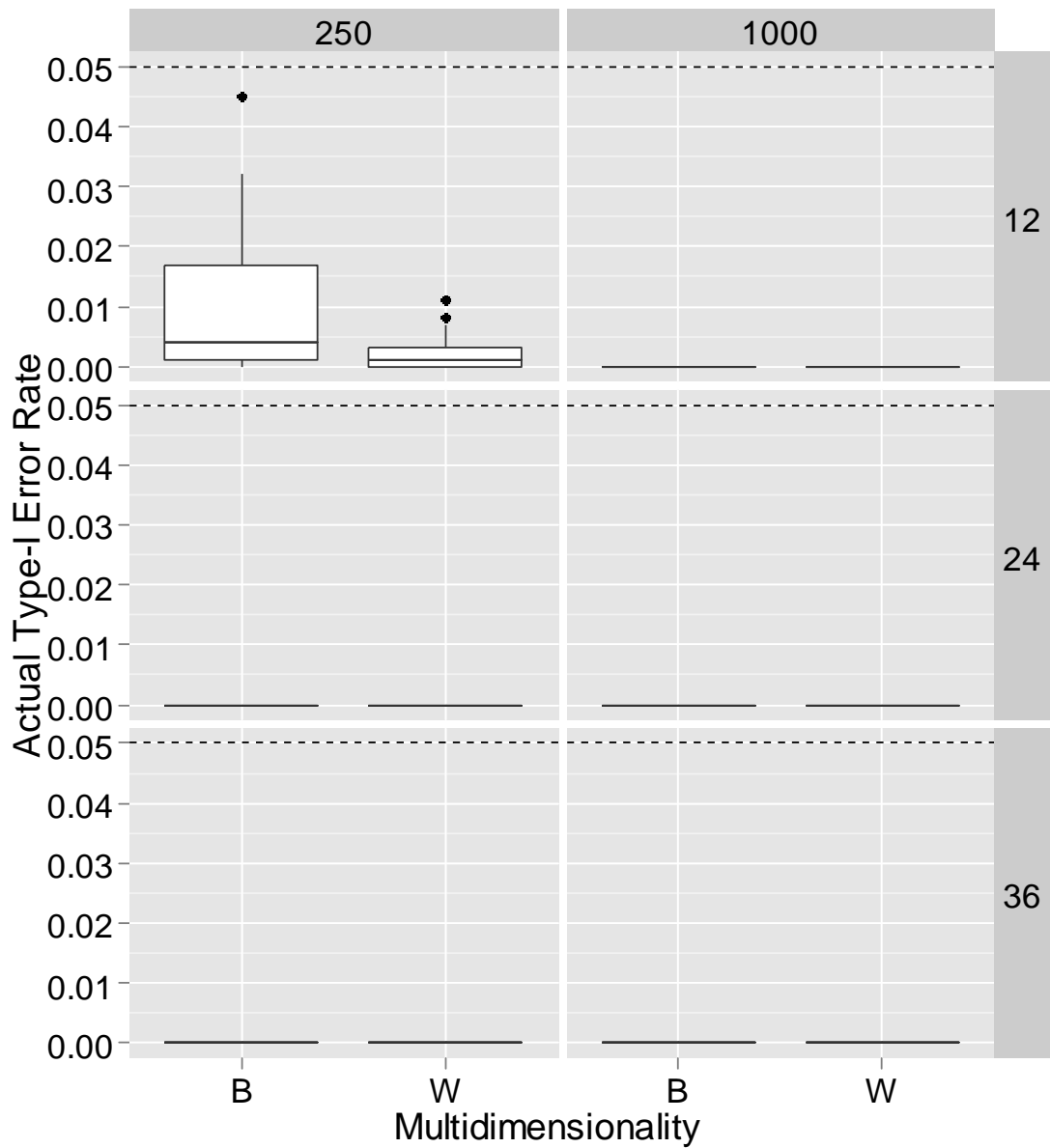


Figure 4.14. Actual Type-I error rates for the RMSEA.

Correctly estimated models are evaluated against the most conservative suggested cut point ($RMSEA = 0.05$). Across multidimensionality, test length (rows), and sample size (columns). The dotted line indicates the nominal significance level.

Considering the actual Type-I error rates of the various item-fit indices reveals patterns of True Model rejection that greatly differ from that observed for the model fit indices. Actual Type-I error rates resulting from the application of the theoretical cut points for the $S-\chi^2$, Modification Index (MI = 3.841), and Wald Test (Wald Test = 3.841) are presented in Figure 4.15 through Figure 4.17, respectively. Unlike the other two item-fit indices, the theoretical cut points for the $S-\chi^2$ are determined for each item separately as a function of the total score point categories containing an appropriate number of observations, therefore, no overall cut point can be stated.

Actual Type-I error rates for the $S-\chi^2$ approximate the nominal significance level for small sample sizes, long tests, and low inter-factor correlations. Decreases in test length, increases in sample size, and shorter test length all contribute to increased Actual Type-I error rates; the median Type-I error rate for 1000 examinees responding to 12 items when latent factors are highly correlated is approximately 0.6. Application of the theoretical cut point to the Modification Index results in approximately nominal Type-I error rates under small sample sizes and high inter-factor correlations. Increases in sample size and decreases in inter-factor correlation result in increased actual Type-I rates; fewer latent factors corresponds to a slight increase in actual Type-I error rates. When two-dimensional models with low inter-factor correlation and 1000 examinees are estimated, the median Type-I error rate is approximately 0.2.

Lastly, the Wald Test under-rejects items estimated as between-item multidimensional with actual Type-I error rates approaching zero. Items that are within-item multidimensional are generally over-rejected under small sample sizes (median Type-I error rates ranging 0.20 to 0.55) and moderately over-rejected under large sample

sizes; high-discrimination and high difficulty correspond to increases in actual Type-I error rates under these conditions.

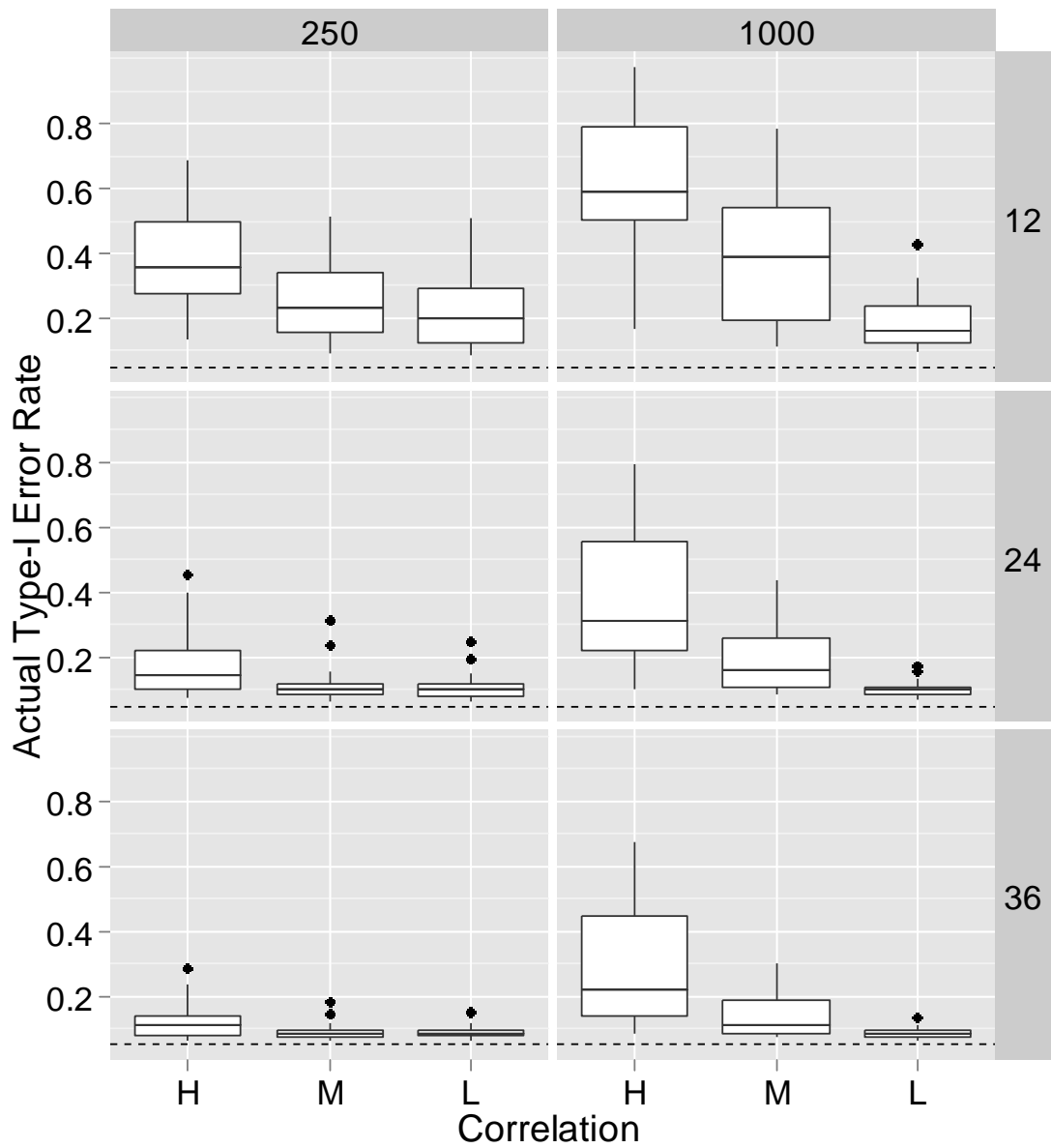


Figure 4.15. Actual Type-I error rates for the $S\text{-}\chi^2$.

Correctly estimated models are evaluated against the theoretical cut point ($MI = 3.841$). Across inter-factor correlation, test length (rows), and sample size (columns). The dotted line indicates the nominal significance level.

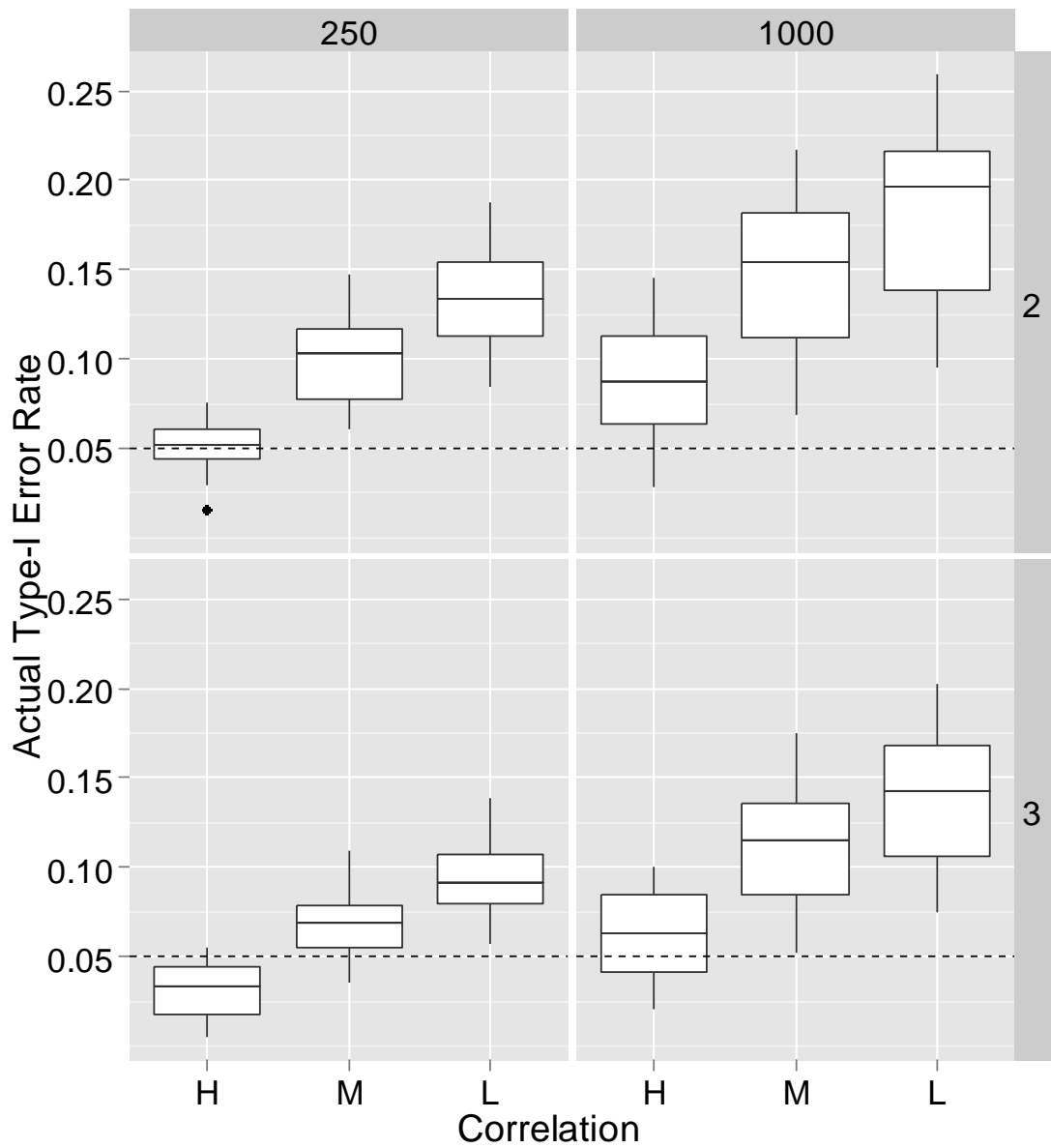


Figure 4.16. Type I error rates for the Modification Index estimated against latent factor 1.

Correctly estimated models are evaluated against the theoretical cut point ($MI = 3.841$). Across inter-factor correlation, number of dimensions (rows), and sample size (columns). The dotted line indicates the nominal significance level.

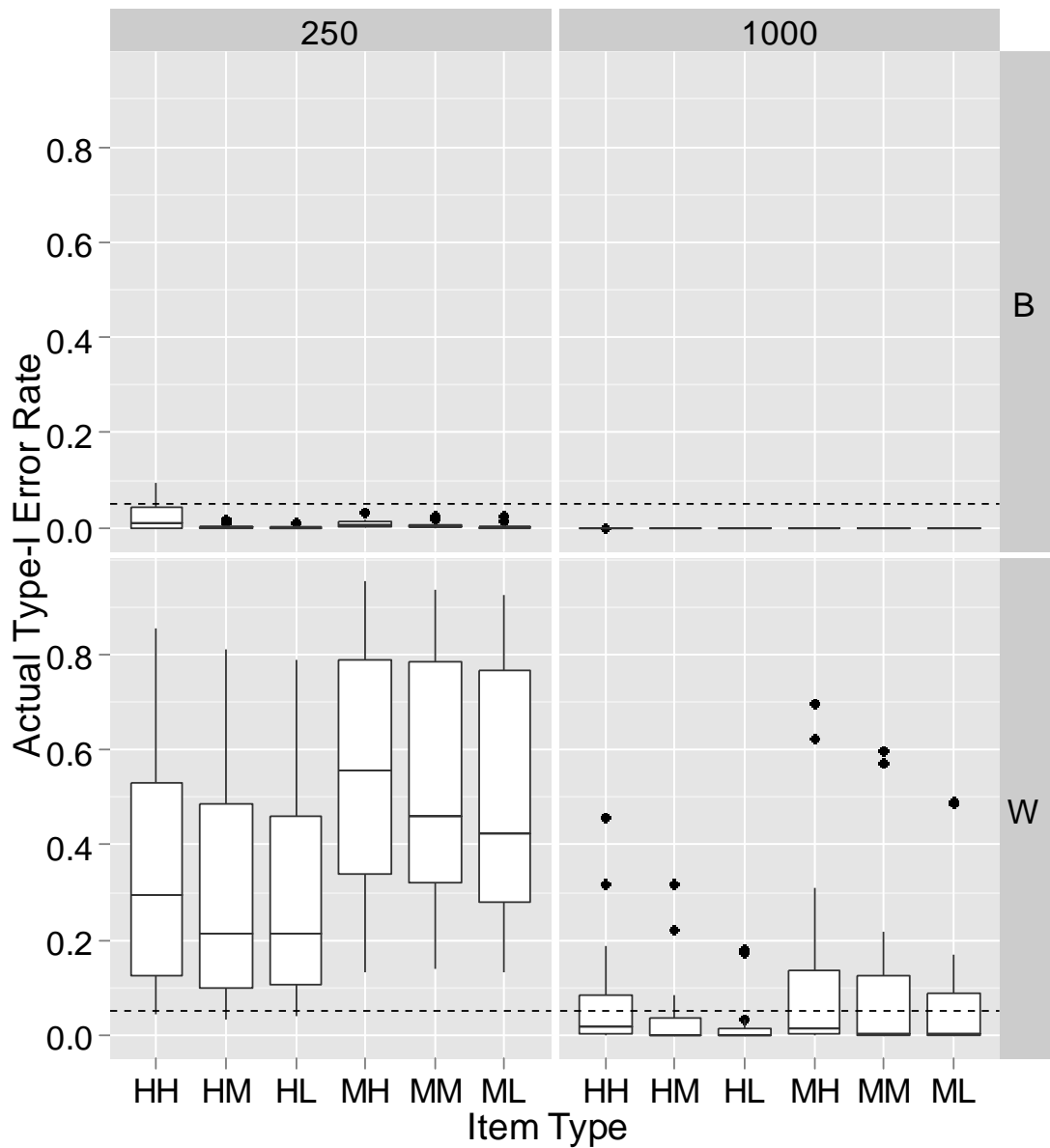


Figure 4.17. Actual Type-I error rates for the Wald Test on latent factor 1.

Correctly estimated models are evaluated against the theoretical cut point (Wald = 3.841). Across item type, multidimensionality (rows), and sample size (columns). The dotted line indicates the nominal significance level.

4.5. Summary

Analysis of the model-fit indices (χ^2/df ratio, RMSEA, and GDDM) and item-fit indices (Modification Index, $S\text{-}\chi^2$, and Wald Test) under true model specification indicate that the 95th percentiles of these statistics, to be subsequently employed as empirically-derived cut points, each demonstrate sensitivity to the various simulation conditions. The current dissertation showed the 95th percentiles of the χ^2/df to be especially sensitive to test length while the RMSEA and GDDM also showed sensitivity to sample size. Previous research by Jackson (2007) found that sample size attributed for 19% of the variance in RMSEA; no studies, however, had so far examined the sensitivity of empirically-derived cut points from a hypothesis-testing perspective. The GDDM demonstrated great sensitivity to item type (i.e., item discrimination and the degree of discrepancy between item difficulty and the mean of the latent factor distributions) even though absolute values of this index remained very small. The three item-fit indices all demonstrated substantial sensitivity to sample size; $S\text{-}\chi^2$ is additionally sensitive to test length, the Modification Index is additionally sensitive to inter-factor correlation, and the Wald Test is additionally sensitive to item multidimensionality.

Given these sensitivities, the use of theoretical or suggested cut points results in actual Type-I error rates that differ greatly from the expected nominal rate of 0.05. For example, the suggested cut points for the χ^2/df and RMSEA fail to reject most models, resulting in underestimated Type-I error rates. This is inconsistent with Marsh, Hau and Wen's (2004) findings for the χ^2/df , for which Type I error rates were 9% and 15% when sample sizes were $n = 250$ and 1,000, likely resulting from differences in parameter specifications. Application of the theoretical cut points to the $S\text{-}\chi^2$ results in inflated

actual Type-I error rates as items are over-rejected in all but a few specific conditions. Actual Type-I error rates for the $S-\chi^2$ were seen to range up to 0.16 and 0.28, increasing with sample size, for models estimated as between- and within-item multidimensional by Zhang and Stone (2008). These results agree with the results of the current study. According to the Modification Indices, items estimated under the True Model are only rejected at the nominal rate for small sample sizes when inter-factor correlation is strong, otherwise actual Type-I error rates are inflated. Finally, the Wald Test is seen to under-reject items that are between-item multidimensional and over-reject items that are within-item multidimensional, though this effect is lessened under large sample sizes.

The results of this section show that the theoretical and suggested cut points are generally inadequate for correctly evaluating model- and item-fit when True Models are estimated under a variety of simulation conditions. Actual Type-I error rates were shown to be both inflated and underestimated depending on the statistic and the specific data generation conditions. It should be noted that suggested cut points, especially those provided by Hu and Bentler (1999), were the result of descriptive analysis of model fit results which attempted to minimize Type-I and Type-II error rates in proposing appropriate, generalized “rule of thumb” criteria. Unlike the Hu and Bentler criteria, the empirical cut points calculated in this study control Type-I error and allow inferential model-fit testing. These cut points are calculated as the 95th percentile resulting from the empirical sampling distribution within each experimental cell, explicitly controlling the nominal significance level as $\alpha = 0.05$.

As stated by Fan, Thompson, and Wang (1999), “the degree of model misspecification should be the major contributor to the variation of a [model- or item-] fit

index” (p. 60); conditions to which a fit statistic demonstrates sensitivity should, therefore, be explicitly considered during model- and item-fit evaluation. Therefore, empirically-derived, design-appropriate cut points are instead employed in subsequent analyses evaluating model and item misspecification in this study. The empirically-derived cut points for each fit index are specified as the 95th percentiles, or 5th percentiles in the case of the Wald Test, resulting from the empirical distribution of 1000 replications within each cell of the simulation design⁵. Utilizing these values thus ensures a nominal Type-I error rate of $\alpha = 0.05$ and precise computations of power given the number of replications in this study.

⁵ Prior simulation work for exploring an appropriate number of replications to help determine these cut-off values with a reasonable degree of precision and without making the running time of the simulation study unduly long, has suggested that 1000 replications is a defensible choice; please see the Appendix for a study exploring this issue.

Chapter 5

Results of Misspecified Model Estimation

The behavior of model- and item-fit indices under correct, or true, model estimation was examined in the previous chapter; the current chapter examines the same indices under the same simulation conditions for moderate or severe model misspecification. First, the bias and precision of item and person parameters is examined in comparison to the results observed for true model estimation. Next, the performance of the model- and item-fit indices is considered in regards to the following research questions:

- 4) How large is the power of different model- and item-fit statistics for detecting different types of Q-matrix misspecification under different test design conditions when the appropriate percentiles from the empirical sampling distribution are used?
- 5) How much of the variation in empirically observed power rates is due to the different Q-matrix misspecification and test design conditions?

Descriptive statistics and power for the model-fit indices are considered first, having applied the empirically-derived cut points calculated from the values obtained under true model estimation. Next, the descriptive statistics and power for the item-fit indices are considered. After addressing these questions, model- and item-fit performance are considered simultaneously, providing holistic information on model evaluation.

5.1. Estimation Issues

Each of the 864 true model conditions were replicated until 250 successful replications for each cell was achieved. Running on a 64-bit dual-core 2.53GHz computer

with 4.00GB of RAM the moderately misspecified conditions took approximately 130 hours to complete and the severely misspecified conditions took approximately 265 hours, for a total of nearly 400 computing hours in estimating and collecting the results of the misspecified models. The majority of the cells in the experimental design required additional replications to achieve the required 250 successful replications; Table 5.1 presents the top 5 simulation conditions for each of 2- and 3-dimensional models requiring additional replications.

Of the 432 moderately misspecified conditions, 260 (60.185%) conditions required a minimum of 251 replications and a maximum of 2425 replications, when estimating models under small sample sizes with 36 high-discrimination / high-difficulty items which follow complex-structure where latent factor are highly correlated. Severely misspecified models required additional replications for 361 (83.565%) of the 432 experimental cells, with a minimum of 251 replications and a maximum of 107,725 replications, when models with 3 weakly correlated factors following simple structure were estimated for 1000 examinees and 12 high-discrimination / moderate difficulty items.

Table 5.1

Top 5 Percentages of Additional Replications Required when Estimated Models are Misspecified

Miss.	Test Length	Sample Size	Multi.	Item Type	2 Dimensions			3 Dimensions		
					L*	M	H	L	M	H
Mod	12	250	W	HH	**	4%	8%			
Mod	12	250	W	HM	4%	5%	8%			
Mod	36	250	W	HH			10%			
Sev	12	250	B	HM	4%	4%		72%	29%	9%
Sev	12	250	W	HM	4%	5%	8%			
Sev	12	250	W	HL			6%			
Sev	12	1000	B	HH				88%	44%	9%
Sev	12	1000	B	HM				431%	70%	9%
Sev	12	1000	B	HL				114%	45%	8%
Sev	12	1000	B	MH	3%					
Sev	12	1000	B	MM	4%	4%				
Sev	12	1000	B	ML				55%	29%	7%

* Indicates inter-factor correlation: L = Low, M = Moderate, and H = High.

** Only the top 5 conditions by inter-factor correlation and number of dimensions are presented for clarity.

Generally, the severely misspecified models required more additional replications than the moderately misspecified models. Additional replications were required for moderately misspecified models when two latent factors were estimated according to complex structure, small sample sizes, and high-discrimination items; when models were severely misspecified, the majority of the models requiring additional replications were comprised of 3 latent factors following simple structure containing 12 items of high discrimination. These results suggest that increasing misspecification results in poor or unreliable estimation, as would be expected. The magnitude of the number of additional replications required in some instances, however, suggests that those conditions are near-unestimable and the results of such models should be interpreted with caution.

Summaries of the root mean-squared error (RMSE) and average bias for MDIFF, MDISC, inter-factor correlations, and ability (i.e., θ) are presented in Table 5.2. Overall, values of the RMSE values for the MDIFF are small (mean = 0.222, median of 0.161) with the largest RMSE values corresponding to the smallest sample size ($n = 250$) but otherwise varied with respect to condition; average bias of MDIFF is also small (mean = -0.001; median = -0.005), indicating that the magnitude of the discrepancy between estimated and generating values is small, with the largest values occurring under the smallest sample size. Recovery of item difficulty is shown to be most dependent on sample size, though the degree of discrepancy is small. Median RMSE and average bias values for the MDIFF parameters are approximately 2 times as large as those seen under true model estimation.

RMSE values for MDISC are slightly larger (mean = 0.332; median = 0.221) and the average bias values are more positive (mean = 0.001; median = 0.003) than those seen for MDIFF, suggesting more discrepancies of greater magnitude. The largest RMSE values are seen for the smallest sample size, the shortest test length, when items are highly discriminating, and factors are highly correlated; average bias shows similar behavior, though values increase as inter-factor correlation becomes stronger. Recovery of discrimination parameters is seen to also be tied to sample size, though also subject to more complex consideration. Median RMSE and average bias values for the MDISC parameters are also approximately 2 times as large as those seen under true model estimation.

Inter-factor correlations across two- and three-dimensional models demonstrate small-to-moderate RMSE values, with means ranging 0.053 to 0.280 and medians of

0.048 to 0.180, where the larger values are associated with two-dimensional models; average bias demonstrates similar ranges and behavior. The largest values of RMSE and relative bias are associated with three-dimensional models demonstrating simple-structure and high inter-factor correlation, with the fewest, highly-discriminating items; the largest average bias values suggest that estimated inter-factor correlations are more than double the generating values. In comparison to the true model, median RMSE values for the inter-factor correlations under misspecified models are 10 times larger and median average bias is up to 3 times larger.

Finally, recovery of examinee ability, θ , is examined. RMSE values are small for ability across two- and three-dimensional models (mean = 0.059 to 0.072; median = 0.065 to 0.070), however, average bias is large (mean = 0.961 to 1.695; median = 0.900 to 0.965), indicating that the majority of the values were recovered within 1 to 2 logits on the θ scale. These statistics were likely influenced by a number of extreme values which were poorly recovered, demonstrated by the wide range of average bias values (-19.045 to 23.364). While bias of approximately 20 is quite large, it is important to note that Mplus does not employ procedures to correct for extreme θ values, unlike IRT software such as Winsteps (Linacre, 2011). While large RMSE values are typically associated with small sample sizes, simple-structure three-dimensional models with highly discriminating and difficulty items, extreme average bias values follow no discernible pattern. Interestingly, the median RMSE and average bias for the latent factors is quite similar to the values seen under true model estimation.

Table 5.2

Descriptive Statistics for RMSE and Average Bias for Moderately Misspecified Models

	Parameter	Min	25th%	Mean	Median	75th%	Max	SD
RMSE	MDIFF	0.071	0.180	0.431	0.291	0.443	19.855	1.154
	MDISC	0.100	0.237	1.506	0.407	2.249	18.498	2.191
	ρ_{12}	0.135	0.344	0.616	0.676	0.907	0.975	0.300
	ρ_{13}	0.141	0.209	0.396	0.405	0.554	0.701	0.170
	ρ_{23}	0.035	0.061	0.079	0.074	0.093	0.158	0.024
	θ_1	0.029	0.038	0.060	0.064	0.072	0.132	0.022
	θ_2	0.030	0.045	0.067	0.067	0.080	0.168	0.026
	θ_3	0.031	0.038	0.061	0.063	0.073	0.126	0.021
Average Bias	MDIFF	-2.551	-0.227	-0.126	0.040	0.097	0.261	0.383
	MDISC	-0.840	-0.188	-0.069	-0.091	0.042	0.621	0.186
	ρ_{12}	0.147	0.405	1.218	0.722	1.640	6.456	1.080
	ρ_{13}	0.177	0.257	1.001	0.702	1.875	4.632	0.808
	ρ_{23}	-0.830	0.508	0.885	0.913	1.351	2.520	0.647
	θ_1	-56.898	0.667	0.962	0.945	1.044	45.734	7.376
	θ_2	-42.839	0.826	2.359	0.972	1.040	99.021	8.679
	θ_3	-33.656	0.721	1.133	0.988	1.263	23.003	5.187

Table 5.3

Descriptive Statistics for RMSE and Average Bias for Severely Misspecified Models

	Parameter	Min	25th%	Mean	Median	75th%	Max	SD
RMSE	MDIFF	0.051	0.195	0.644	0.319	0.585	31.832	2.315
	MDISC	0.144	0.281	2.457	0.497	4.135	26.081	3.492
	ρ_{12}	0.189	0.438	0.691	0.750	0.956	0.991	0.285
	ρ_{13}	0.172	0.228	0.444	0.453	0.629	0.740	0.190
	ρ_{23}	0.047	0.075	0.122	0.099	0.172	0.265	0.058
	θ_1	0.027	0.037	0.058	0.063	0.071	0.124	0.020
	θ_2	0.030	0.040	0.063	0.066	0.075	0.172	0.024
	θ_3	0.029	0.039	0.061	0.064	0.073	0.126	0.022
Average Bias	MDIFF	-4.725	-0.627	-0.328	0.021	0.094	0.826	0.723
	MDISC	-1.120	-0.205	-0.033	-0.069	0.175	0.577	0.254
	ρ_{12}	0.196	0.466	1.152	0.938	1.569	3.691	0.877
	ρ_{13}	0.199	0.284	0.969	0.715	1.419	2.752	0.761
	ρ_{23}	-1.541	0.902	1.095	1.119	1.323	2.803	0.543
	θ_1	-49.678	0.631	1.300	0.940	1.098	56.566	8.595
	θ_2	-45.735	0.797	1.865	0.966	1.094	37.962	6.192
	θ_3	-22.727	0.797	1.508	0.979	1.091	32.819	6.142

Overall, variability of parameter recovery as described by RMSE appears to be small and impacted mainly by sample size, suggesting that parameters are less precise at the smallest sample size. The magnitude of the discrepancies, indicated by average bias, is generally small for item parameters but suggests the presence of overestimated values, in the case of inter-factor correlations, and extreme values, for ability estimates, frequently associated with three-dimensional models following simple-structure with highly-discriminating items. These values are typically increased over those seen under true model estimation, indicating the effect of misspecification. While mean and median values of the latent factors are recovered similar to values under true model estimation, though true models demonstrated a narrower range of values, the inter-factor correlations appear to be the least well-recovered parameters suggesting that these parameters are more susceptible to model misspecification.

5.2. Analysis of Model-Fit Indices under Model Misspecification

5.2.1 Distributional Characteristics of Model Fit Indices

Moderately and severely misspecified models were estimated and the values of the χ^2/df ratio, RMSEA, and GDDM model-fit indices submitted to separate ANOVAs including test design and model conditions as factors. Descriptive statistics for these indices under model misspecification are presented graphically in Figure 5.1 to Figure 5.3 and are summarized according to the simulation conditions for which the specific index demonstrates the greatest sensitivity resulting from the factorial ANOVA. This means that the ranges presented in the tables and figures represent ranges of the fit index values across the simulation conditions which are not presented. Table 5.4 presents the

percentages of variance associated with main effects and interactions thereof for which the model-fit indices demonstrated sensitivity ($\eta^2 \geq 1.000$).

Table 5.4

Selected Percentages of Variance for Model-Fit Indices Under Model Misspecification, by Simulation Conditions

Source	χ^2/df	RMSEA	GDDM
Model Misspecification (0)	0.105	0.212	<u>1.607</u>
Number of Dimensions (1)	<u>5.828</u>	<u>10.918</u>	<u>17.275</u>
Test Length (2)	<u>3.600</u>	<u>6.712</u>	<u>5.289</u>
Sample Size (3)	<u>23.943</u>	<u>1.648</u>	<u>2.942</u>
Item Multidimensionality (4)	0.800	<u>1.257</u>	0.137
Inter-Factor Correlation (5)	<u>22.070</u>	<u>52.471</u>	<u>24.827</u>
Item Type (6)	<u>6.237</u>	<u>9.851</u>	<u>18.634</u>
0*4	0.003	0.011	<u>1.314</u>
0*5	0.045	0.038	<u>1.072</u>
1*3	<u>2.385</u>	0.011	0.012
1*4	<u>1.663</u>	<u>2.606</u>	0.075
1*5	<u>2.232</u>	<u>1.268</u>	<u>2.534</u>
1*6	0.822	0.538	<u>3.321</u>
2*3	<u>1.130</u>	0.081	0.497
2*5	<u>1.583</u>	<u>1.155</u>	0.035
2*6	0.990	<u>1.052</u>	<u>1.089</u>
3*5	<u>9.537</u>	0.090	0.119
3*6	<u>2.827</u>	0.158	0.191
5*6	<u>2.501</u>	<u>1.422</u>	<u>2.502</u>
3*5*6	<u>1.089</u>	0.107	0.056
Residuals	1.090	2.198	7.819

Note: Highlighted cells indicate conditions presented in the box-and-whiskers plots.

5.2.1.1 Results for χ^2/df

It is first notable that the majority (almost 99%) of the variance in the χ^2/df ratio is attributable to main effects and interactions of the simulation conditions. Of the conditions demonstrating sensitivity under model misspecification, the greatest among these are the main effects of sample size ($\eta^2 = 23.943\%$) and inter-factor correlation ($\eta^2 = 22.070\%$) and the first-order interaction of these two factors ($\eta^2 = 9.537\%$). Item type is attributable for the next largest percentage of variance ($\eta^2 = 6.237\%$) while the

χ^2/df ratio is shown to be insensitive to degree of model misspecification ($\eta^2 = 0.105\%$) or multidimensionality ($\eta^2 = 0.800\%$), representing model estimation by different types of Q-matrices. These conditions differ from the conditions demonstrating sensitivity under true model estimation, which included test length and the interaction of sample size and item type. The effect of these sensitivities is presented as a 90%-winsorized box-and-whiskers plot in Figure 5.1 according to sample size, inter-factor correlation, and item type. Values of the χ^2/df approximate 1.0 under small sample sizes with high inter-factor correlations, suggesting that the misspecified models fit the data, and increase with sample size and item discrimination while decreasing with inter-factor correlation and item difficulty; the effect of item type becomes more pronounced as inter-factor correlation decreases. Values of the χ^2/df are largest, indicating the model misfit, for large sample sizes, low inter-factor correlations, and items of high-discrimination / low-difficulty – resulting in an inter-quartile range (IQR) of $\chi^2/\text{df} = [8.606, 16.591]$ and a maximum of $\chi^2/\text{df} = 37.884$. Descriptive statistics are presented in the Appendix according to the same conditions as the box-and-whiskers plot.

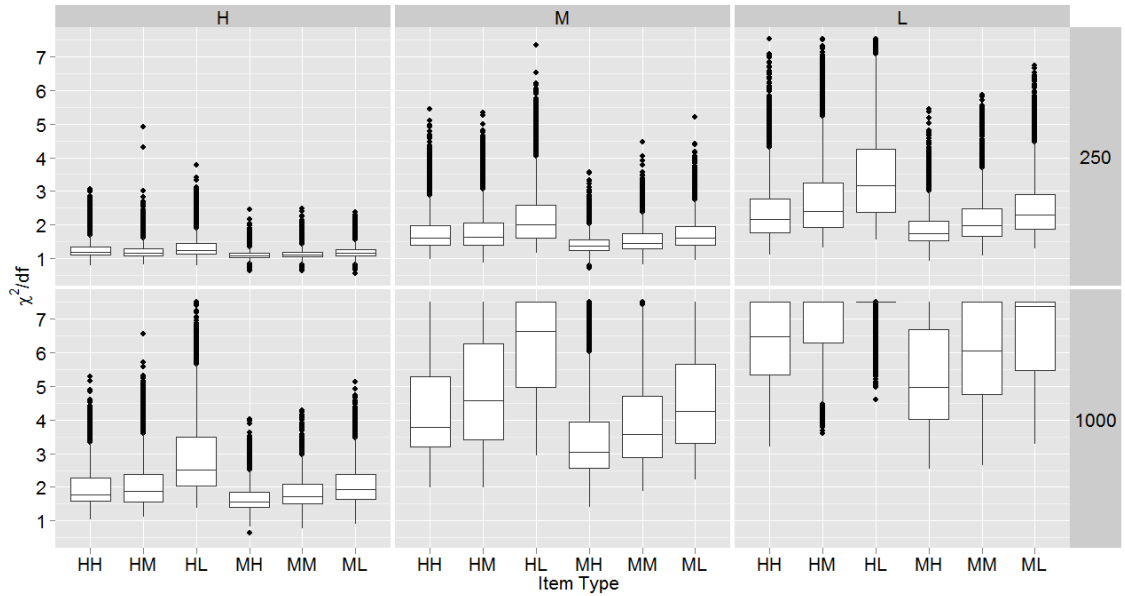


Figure 5.1. Box-and-Whiskers Plots for χ^2/df under Model Misspecification.

Presented according to conditions associated with sensitivity: item type (HH = high discrimination / high difficulty; HM = high discrimination / moderate difficulty; HL = high discrimination / low difficulty; MH = moderate discrimination / high difficulty; MM = moderate discrimination / moderate difficulty; ML = moderate discrimination / low difficulty), sample size (rows), and inter-factor correlation (columns; H = correlations of 0.75; M = correlations of 0.50; L = correlations of 0.25).

5.2.1.2 Results for RMSEA

Based on the model-fit χ^2 , the RMSEA demonstrates sensitivities similar to the χ^2/df ratio – almost 98% of the variance in the RMSEA is attributable to main effects and interactions of the simulation conditions. Similar to the χ^2/df , the RMSEA demonstrates sensitivity to inter-factor correlation ($\eta^2 = 52.471\%$) and item type ($\eta^2 = 9.851\%$); unlike the χ^2/df , the number of dimensions ($\eta^2 = 10.918\%$) is included in the top three simulation conditions for sensitivity as resulting from the factorial ANOVA. This is quite different from the conditions demonstrating sensitivity under true model estimation (i.e., test length, sample size, and multidimensionality). The 90%-winsorized box-and-whiskers plot for the RMSEA is presented in Figure 5.2 according to inter-factor correlation, item type, and number of dimensions and the corresponding descriptive statistics are included in the Appendix.

Similar to the χ^2/df , model-fit index, RMSEA values are seen to decrease with inter-factor correlation, increase with item discrimination, and decrease with item difficulty when misspecified models are estimated. Additionally, RMSEA values decrease with the number of dimensions or latent factors. The lowest RMSEA values resulting from misspecified models are found when highly-correlated 3-dimensional models with moderately-discriminating / high-difficulty items are estimated (IQR = [0.015, 0.023], maximum RMSEA = 0.062) while the largest RMSEA values result from weakly-correlated 2-dimensional models comprised of highly-discriminating / low-difficulty items (IQR = [0.100, 0.138], maximum RMSEA = 0.203).

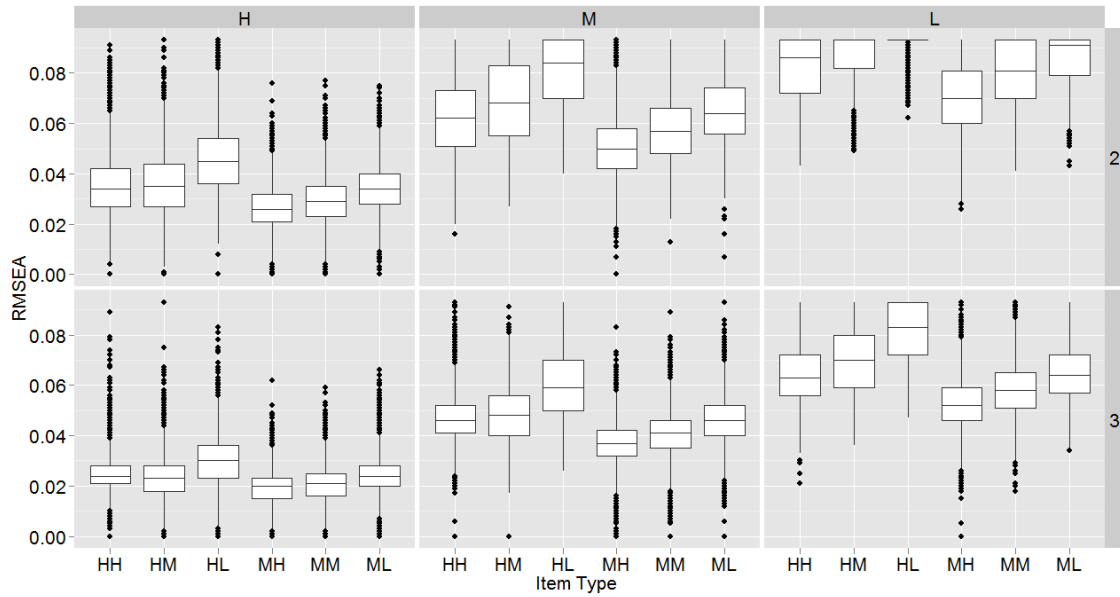


Figure 5.2. Box-and-Whiskers Plots for RMSEA under Model Misspecification.

Presented according to conditions associated with sensitivity: item type, number of latent factors(rows), and inter-factor correlation (columns).

5.2.1.3 Results for GDDM

Lastly, when estimated models are misspecified the GDDM demonstrates less sensitivity to simulation conditions than the other model-fit indices; 92.181% of variance is attributable to the simulation conditions. The GDDM demonstrates greatest sensitivity

to inter-factor correlation ($\eta^2 = 24.827\%$), then item type ($\eta^2 = 18.634\%$), and number of dimensions

($\eta^2 = 17.275\%$). These simulation conditions are included as factors in the presentation of the descriptive statistics (Appendix) and the 90%-winsorized box-and-whiskers plot (Figure 5.3). The effect of sensitivity to item type is very similar to that seen under true model estimation. The best-fitting misspecified models (3 highly-correlated dimensions estimated for highly-discriminating / high-difficulty items) demonstrate GDDM values with $\text{IQR} = [0.004, 0.005]$ and the maximum value is $\text{GDDM} = 0.007$ while the worst-fitting models (2 weakly-correlated dimensions estimated for moderately-discriminating / low-difficulty items) demonstrate GDDM values with $\text{IQR} = [0.012, 0.170]$ with a maximum value of $\text{GDDM} = 0.027$.

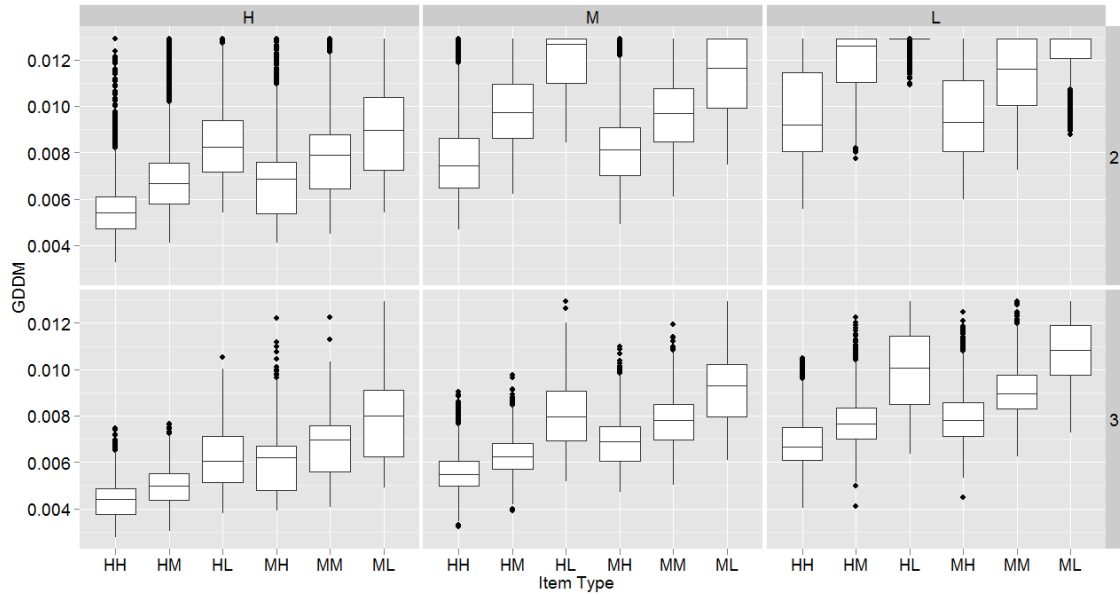


Figure 5.3. Box-and-Whiskers Plots for GDDM under Model Misspecification.

Presented according to conditions associated with sensitivity: item type, number of dimension (rows), and inter-factor correlation (columns).

5.2.2 Power of Model-fit Indices

All of the models estimated in the moderate and severe misspecification conditions were, obviously, misspecified to a degree via alternate-factoring or under-factoring of specific elements of the estimating Q-matrix. As such, power can be calculated as the average rate of model rejection, aggregated over simulation conditions and across replications. Specifically, values of the χ^2/df ratio, RMSEA, and GDDM model-fit indices are compared to suggested or empirical cut points and subsequently indicating model fit or misfit.

The empirically-determined cut points were determined separately for each cell of the simulation design as the 95th percentile values for all model-fit indices, thereby fixing the nominal Type-I error rate to approximately 0.05. This is an approximate rate because 1000 replications of the true models still results in some small imprecision at determining an exact cut-off point to achieve the exact nominal rate even though the preliminary work showed that the approximation is reasonably close (see the Appendix).

Even though the χ^2/df ratio, RMSEA, and GDDM demonstrate wide variation in values resulting from the various simulation conditions under moderate and severe model misspecification, the statistics generally demonstrate moderate to high power in correctly rejecting misspecified models. Specific sensitivities for the power of each of the model-fit indices to the various simulation conditions are presented in Table 5.5 as the results of yet another factorial ANOVA.

Table 5.5
Selected Percentages of Variance for Power of Model-Fit Statistics

Source	χ^2/df	RMSEA	GDDM
Model Misspecification (0)	0.067	0.072	0.104
Number of Dimensions (1)	<u>3.441</u>	<u>3.341</u>	<u>3.288</u>
Test Length (2)	<u>6.007*</u>	<u>6.168</u>	<u>4.898</u>
Sample Size (3)	<u>10.458</u>	<u>10.234</u>	<u>9.478</u>
Item Multidimensionality (4)	0.126	0.119	0.043
Inter-Factor Correlation (5)	<u>17.769</u>	<u>17.195</u>	<u>17.969</u>
Item Type (6)	<u>2.324</u>	<u>2.268</u>	<u>4.680</u>
1*3	<u>2.547</u>	<u>2.531</u>	<u>2.157</u>
1*5	<u>3.848</u>	<u>3.781</u>	<u>4.467</u>
2*3	<u>4.312</u>	<u>4.589</u>	<u>3.094</u>
3*6	<u>1.456</u>	<u>1.484</u>	<u>2.922</u>
5*6	<u>2.456</u>	<u>2.385</u>	<u>5.263</u>
3*5	<u>14.294</u>	<u>14.074</u>	<u>13.229</u>
2*5	<u>6.488</u>	<u>6.829</u>	<u>5.504</u>
2*3*5	<u>4.111</u>	<u>4.572</u>	<u>2.932</u>
3*5*6	<u>1.317</u>	<u>1.350</u>	<u>2.880</u>
1*3*5	<u>2.562</u>	<u>2.605</u>	<u>2.657</u>
Residuals	3.435	3.555	2.679

* Cells highlighted in dark grey indicate top-three sources of variance; cells highlighted in light grey indicate top main effects suggested by top-three interactions.

5.2.2.1 Power for χ^2/df

Figure 5.4 presents the power values and ranges for the χ^2/df model-fit statistic, summarized according to the simulation conditions that for which power of the χ^2/df was shown to be mostly sensitive: sample size, test length, and inter-factor correlation. Across conditions, the χ^2/df demonstrates ranges of power that approach 1.0, however, when sample size is small, the test is comprised of few items, and when inter-factor correlations are strong the summarized simulation conditions result in ranges of power less than 1.0. That is to say, the ability of the χ^2/df ratio to correctly reject misspecified models improves as sample size and test length increase, regardless of other conditions such as those included in the current study. Specifically, for short tests with a small sample size and highly-correlated latent factors, the power of the χ^2/df demonstrates an

IQR = [0.271, 0.714] with a median of 0.432. When the empirically-derived cut points are applied to misspecified models with large sample sizes, many items, and low inter-factor correlations, however, all of the models are correctly rejected.

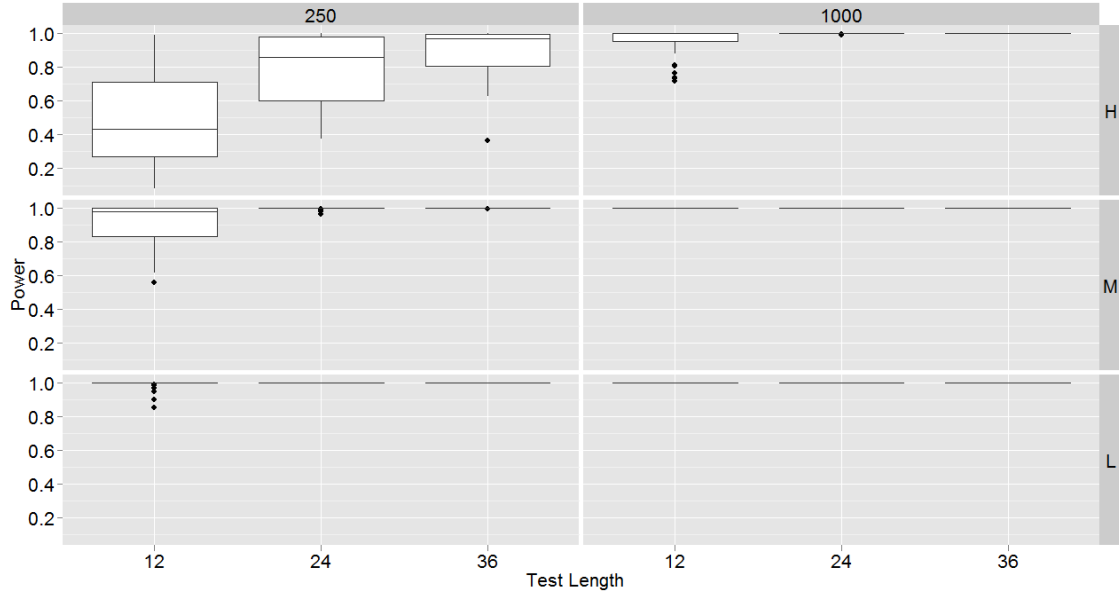


Figure 5.4. Box-and-Whiskers Plots for Power of χ^2/df ratio.

Presented according to conditions associated with sensitivity: test length, inter-factor correlation (rows), and sample size (columns).

5.2.2.2 Power for RMSEA

The RMSEA demonstrates a pattern similar to that of the χ^2/df ratio with power that approaches 1.0 as sample size and test length increase and when inter-factor correlation is weak (Figure 5.5). Again, the lowest and widest proportions of correctly rejected misspecified models occurred for those models comprised of 12 items, 250 examinees, and highly-correlated latent factors (IQR = 0.285 to 0.721; median = 0.454).

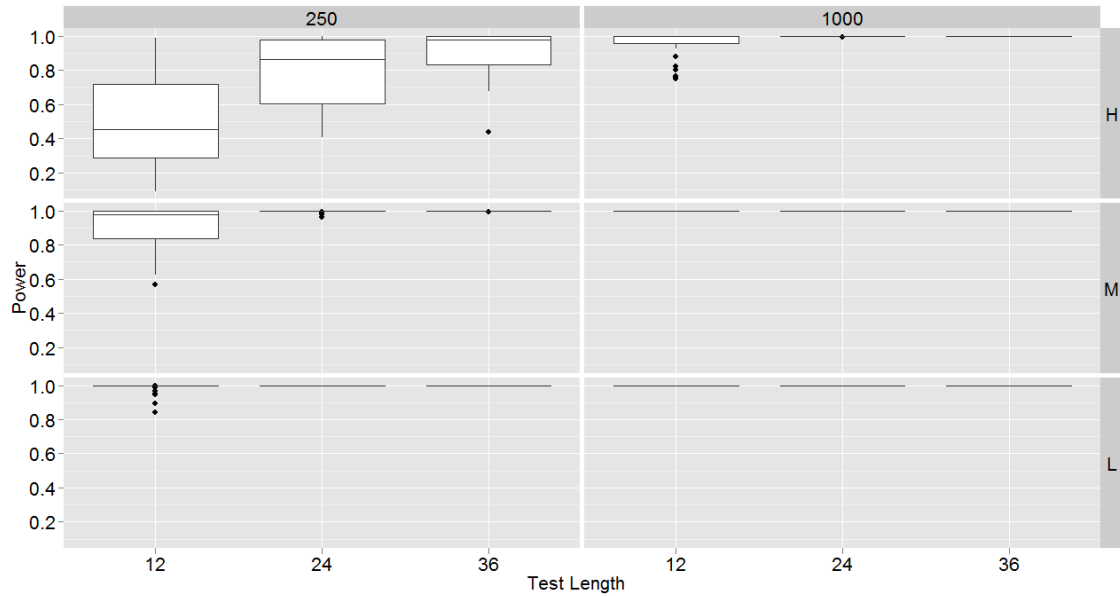


Figure 5.5. Box-and-Whiskers Plots for Power of the RMSEA.

Presented according to conditions associated with sensitivity: test length, inter-factor correlation (rows), and sample size (columns).

5.2.2.3 Power for GDDM

Box-and-whisker plots illustrating power for the GDDM to correctly reject misspecified models is presented in Figure 5.6 and, like the other model-fit indices, shows moderate-to-high power across simulation conditions, including inter-factor correlation, sample size, and test length. Misspecified models are most often correctly rejected when weakly-correlated factors are estimated for long tests and large sample sizes; $IQR = [1.000, 1.000]$, median = 1.000. Conversely, power is worst for small sample sizes when models with highly-correlated latent factors are estimated from short tests; $IQR = [0.384, 0.797]$, median = 0.620.

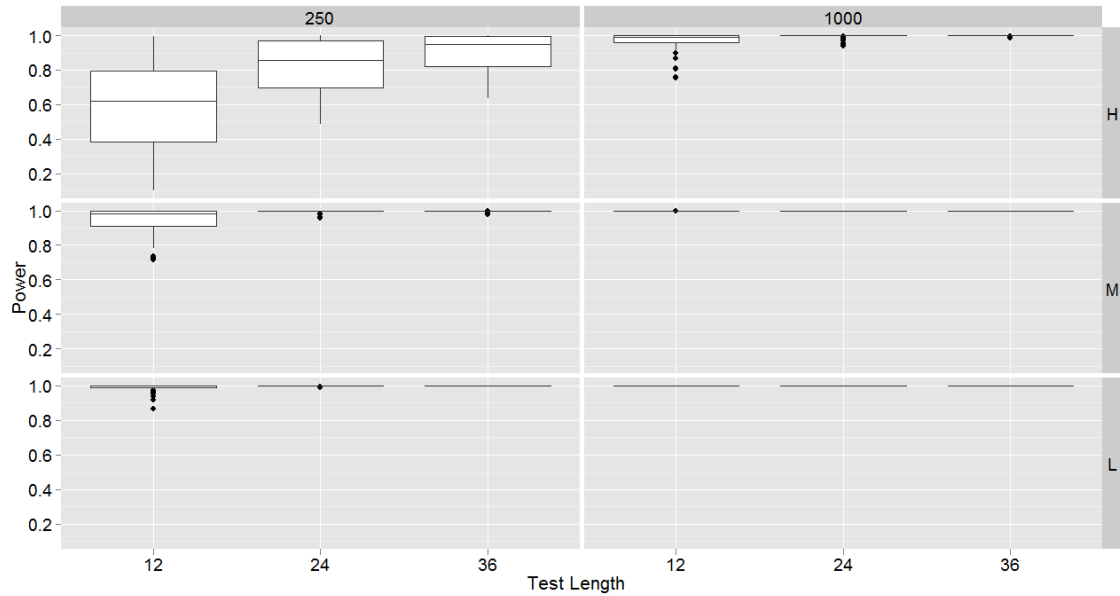


Figure 5.6. Box-and-Whiskers Plots for Power of the GDDM.

Presented according to conditions associated with sensitivity: test length, inter-factor correlation (rows), and sample size (columns).

5.2.3 Summary for Model-Fit Indices

Overall, the model-fit indices are shown to demonstrate large values, indicating misfit of the misspecified models, when items are highly discriminating, tests are short in length, and sample sizes are small. Following from this, the highest power rates for correctly rejecting misspecified models very clearly correspond to large sample sizes, long test lengths, and weakly-correlated latent factors. Conversely, the model-fit indices perform poorly in rejecting the misspecified models when sample sizes are small, tests are short, and the dimensions are highly-correlated. The conditions under which models were seen to demonstrate greatest estimation difficulties (severely misspecified 3-dimensional models with weakly- and moderately-correlated factors following simple, comprised of high-discrimination items) do correspond with conditions of poor fit but this does not appear to be a determining factor in the power associated with the model-fit indices.

As expected, the χ^2/df and RMSEA demonstrate similar patterns of sensitivity to simulation conditions and power when rejecting misspecified models; both fit indices correctly reject misspecified models at rates approaching 1.0. Further, these indices best detect misfit when latent factors are distinct (i.e., low inter-factor correlation) and items target the distribution of the latent factors (i.e., low difficulty, interpreted as minimal discrepancy from the examinee latent variable distribution). The seminal research by Hu and Bentler (1999) showed that power of the RMSEA increased with both degree of misspecification and sample size; at RMSEA = 0.045, the cut point closest to the mean and median of the empirical cut points in this dissertation, Hu and Bentler reported power that approached 1.0 for the RMSEA. Jackson (2007) reported the power of the ML- χ^2 to increase with sample size, test length, and magnitude of factor loadings. For the smallest misspecification and $n = 200$ power was shown to range 0.13 to 0.26 but approaches 1.00 when sample size was increased to $n = 800$. Factor loadings employed in the Jackson (2007) study ranged 0.60 to 0.80, which yield MDISC values lower than those simulated in the current study. The results of this dissertation follow those presented in previous research with respect to the χ^2/df and RMSEA model-fit indices.

Though it is not based off of the model χ^2 , the GDDM demonstrates magnitudes and patterns of power rates similar to the other model-fit indices. For sample sizes of 1000, the GDDM almost perfectly rejects all of the misspecified models; power rates are lower under smaller sample sizes though still moderate-to-high. The study by Levy and Svetina (2010) found that the GDDM correctly identified model misspecification when a more restrictive model was estimated for 1000 examinee responses to a 36-item test with uncorrelated latent factors. When a 2-factor simple structure model was estimated,

identification rates for the GDDM approached 1.00 for data generated according to 3 uncorrelated latent factors following complex-structure and decreased to 0.08 for data generated according to 2 latent factors correlated at $\rho = 0.5$ following complex-structure. These results generally agree with those found in the current study, as power is seen to be strongly influenced by the degree of inter-factor correlation. Though not one of the top sources of sensitivity, it is important to note that power under the GDDM is influenced by item type – an effect which strongly appeared under true model estimation.

When estimating misspecified models, it is important to highlight that neither the χ^2/df , RMSEA, or the GDDM demonstrated sensitivity the degree of misspecification (moderate versus severe) and only the GDDM demonstrated sensitivity to multidimensionality, which reflects different types of Q-matrices. Additionally, all of the model-fit indices demonstrated sensitivity to item type (to some degree).

5.3. Analysis of Item-Fit Indices

5.3.1 Distributional Characteristics of Item-Fit Indices under Model Misspecification

In addition to the typical simulation conditions, type of misspecification is added to the following analyses of the item-fit indices indicating that items were (1) correctly estimated, (2) alternate-factor misspecified, or (3) underfactored, as described in Chapter 3. Prior to the analysis the item-fit indices, the effect of alternate-factoring on the estimated MDISC values is explored to ensure that this misspecification does not simply result in the deletion of factor loading and the addition of a “nuisance” parameter – a relatively insignificant factor loading or MDISC value.

When items were misspecified according to alternate-factoring, the effect of deleting the primary factor while adding a nuisance factor can be assessed by examining

the RMSE and bias of the resulting MDISC estimates on the misspecified factor. It can be expected that a nuisance factor would be indicated by low estimated MDISC values (i.e., weak factor loadings) that differ greatly from the original MDISC values and likely approach zero. Therefore, RMSE values would be expected to be large and bias values would be negative, indicating smaller estimates of MDISC compared to the generating values. Further, key descriptive statistics for alternate-factored items can be calculated to ascertain whether these parameters suggest the presence of a nuisance factor. Table 5.6 contrasts descriptive statistics for MDISC values resulting from correct and alternate-factored items as well as presenting the ratio of those values and the RMSE and bias, aggregated over all other conditions. These results are also presented graphically in Figure 5.7.

As can be seen from these results, the MDISC values associated with items that have been misspecified due to alternate-factoring are systematically lower than values for the correctly specified items; alternate-factored values for MDISC are between 0.637 and 1.424 while the MDISC values for correctly specified items are between 0.813 and 2.183. This is further indicated by the generally negative bias values associated with the alternate-factored items. The RMSE, however, is notably smaller for these misspecified items than for the correctly specified items. From these results, it appears that items misspecified according to alternate-factoring yield lower MDISC parameter estimates which are not small enough to be considered nuisance parameters – the MDISC values are substantial in comparison to the correctly estimated parameter values.

Table 5.6
Descriptive Statistics for MDISC Values when Items were Correctly Specified or Alternate-Factored

Statistic	Misspecification	Mean	Median	SD	Min	Max
Mean	Correct	1.260	1.209	0.286	0.813	2.183
	Alternate	0.978	0.937	0.217	0.637	1.424
Ratio	Alternate / Correct	0.780	0.790	0.079	0.530	0.927
RMSE	Correct	1.314	1.116	1.032	0.162	6.608
	Alternate	0.312	0.279	0.143	0.096	0.993
Average Bias	Correct	-0.149	-0.136	0.102	-0.896	0.004
	Alternate	-0.307	-0.260	0.242	-2.902	-0.076

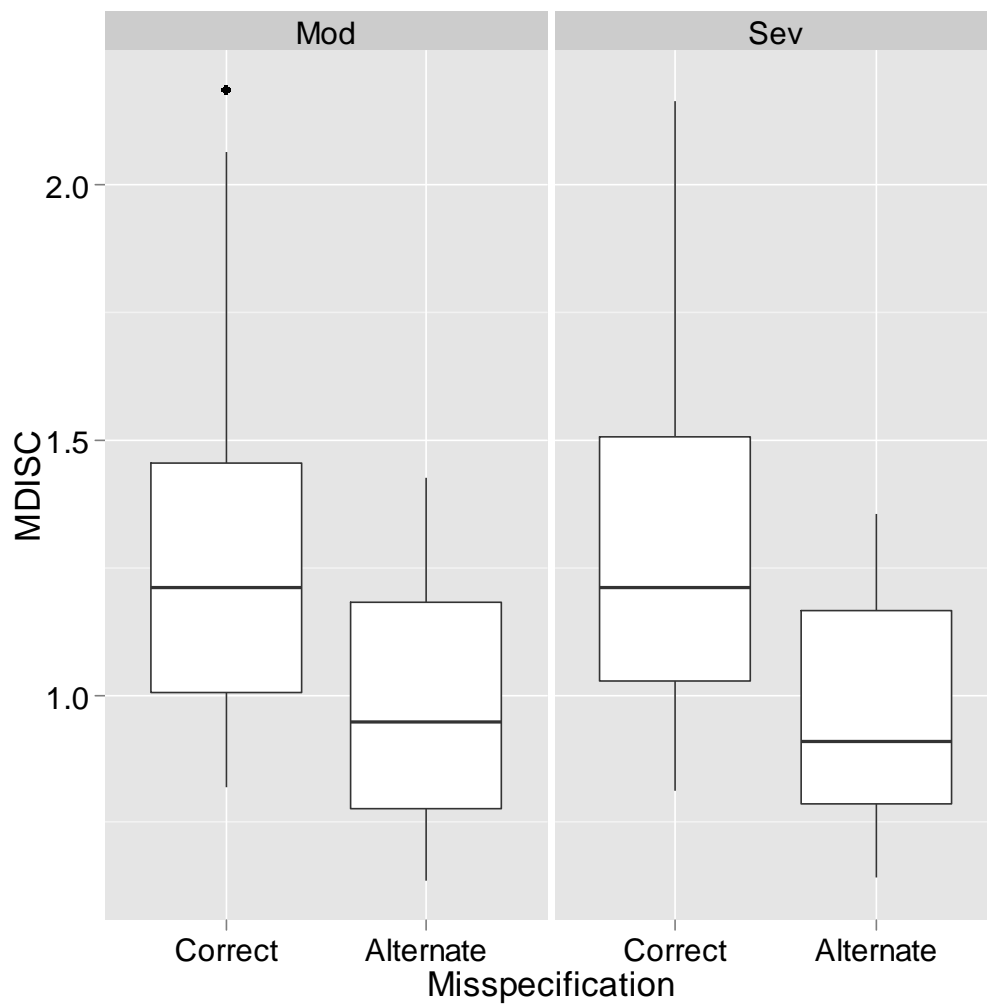


Figure 5.7. Box-and-Whiskers Plots for MDISC Values when Items were Correctly Specified or Alternate-Factored.

Misspecification of items according to alternate-factoring or underfactoring does not occur across all levels of all the other simulation conditions, resulting in an incomplete factorial design (Table 5.7). Such a design compromises the use of ANOVA in calculating sensitivity of the item-fit indices as the sum-of-squares are no longer orthogonal. A full-factorial design for analyzing item-fit values is achieved through the creation of a compound factor comprised of model misspecification, estimated item multidimensionality, and type of item misspecification. This compound factor is included as a simulation design condition in subsequent analysis of the item-fit statistics.

Table 5.7

Types of Item Misspecification Present by Model Misspecification and Item Multidimensionality

Model Misspecification	Estimated Item Multidim.	Type of Misspecification		
		Correct	Alternate-Factoring	Under-factoring
Moderate	Between	x	x	
	Within	x		
Severe	Between	x	x	x
	Within	x		

Descriptive statistics for the $S-\chi^2$, Modification Index, and Wald Test item-fit indices under model misspecification are available in the Appendix and presented graphically in Figure 5.8 to Figure 5.10 according to main effects and interactions demonstrating sensitivity in the item-fit indices. For each main effect and interaction demonstrating sensitivity ($\eta^2 \geq 1.000$), Table 5.8 presents the percentages of variance associated with simulation conditions and their interactions resulting from the factorial ANOVAs conducted for each item-fit index.

Table 5.8

Selected Percentages of Variance for Item-Fit Statistics by Simulation Condition Under Model Misspecification

Source	S- χ^2	Modification Index			Wald Test		
		1	2	3	1	2	3
Number of Dimensions (1)	0.245*	<u>3.035</u>	<u>4.353</u>		<u>1.233</u>	<u>5.342</u>	
Test Length (2)	0.100	0.079	0.140	0.126	<u>1.244</u>	<u>1.348</u>	<u>2.106</u>
Sample Size (3)	0.415	<u>4.332</u>	<u>2.937</u>	<u>3.831</u>	<u>18.893</u>	<u>21.407</u>	<u>24.159</u>
Inter-factor Correlation (5)	0.261	<u>3.536</u>	<u>2.405</u>	<u>3.100</u>	<u>1.015</u>	0.362	<u>2.813</u>
Item Type (6)	0.157	0.823	0.379	<u>1.900</u>	<u>10.091</u>	<u>13.132</u>	<u>11.747</u>
Misspecification Type (7)	<u>2.917</u>	<u>1.104</u>	0.923	<u>1.512</u>	<u>40.534</u>	<u>36.090</u>	<u>32.334</u>
1*2	0.031	0.037	0.034		0.001	0.177	
1*3	0.011	<u>1.492</u>	<u>2.134</u>		0.118	0.564	
1*5	0.102	<u>1.199</u>	<u>1.695</u>		0.014	0.120	
1*7	0.730	<u>2.446</u>	<u>2.154</u>		0.989	0.023	
3*5	0.015	<u>1.599</u>	<u>1.043</u>	<u>1.307</u>	0.099	0.031	0.261
3*6	0.046	0.418	0.192	<u>1.014</u>	0.974	<u>1.266</u>	<u>1.137</u>
3*7	0.198	0.554	0.435	0.753	<u>4.010</u>	<u>3.146</u>	<u>3.111</u>
5*7	<u>1.349</u>	0.434	0.366	0.604	0.703	0.563	<u>1.620</u>
6*7	0.834	0.108	0.097	0.529	<u>2.574</u>	<u>2.988</u>	<u>3.248</u>
1*2*7	0.213	<u>1.327</u>	0.211		0.040	0.015	
1*3*7	0.226	<u>1.243</u>	0.997		0.081	0.010	
1*6*7	<u>1.024</u>	0.284	0.154		0.328	0.012	
Residuals	78.047	68.911	74.432	82.054	15.391	11.840	15.442

* Cells highlighted in dark grey indicate top-three sources of variance; cells highlighted in light grey indicate top main effects suggested by top-three interactions.

5.3.1.1 Distributional Characteristics of the $S-\chi^2$

The $S-\chi^2$ demonstrates sensitivity to simulation conditions though the majority of variance in this item-fit statistic is due to unique item variability ($\eta^2 = 78.047\%$). Specifically, the $S-\chi^2$ shows sensitivity to type of misspecification ($\eta^2 = 2.917\%$), the first-order interaction of inter-factor correlation and type of misspecification ($\eta^2 = 1.349\%$), and the second-order interaction of number of dimensions with item type and misspecification type ($\eta^2 = 1.024\%$). These sensitivities differ from those observed under true model estimation (i.e., test length, sample size, and inter-factor correlation). The effect of these sensitivities is presented in Figure 5.8 as a 90%-winsorized box-and-whiskers plot according to the main effects suggested by the sensitivity analysis: number of dimensions, inter-factor correlation, and type of misspecification. The interaction of misspecification type with inter-factor correlation is apparent as values of the $S-\chi^2$ generally increase with inter-factor correlation (suggesting misfit) and type of misspecification (alternate-factoring and underfactoring) and decrease with item multidimensionality. Interestingly, values of the $S-\chi^2$ appear to decrease slightly across degree of model misspecification, indicated via the type of misspecification factor. Fit is worst (i.e., largest values) when 2-dimensional models were estimated as moderately misspecified and between-item multidimensional items were estimated as associated with an alternate factor: IQR = 18.680, 45.400]; median = 28.258. Alternately, the best fit occurs for between-item multidimensional items correctly estimated within weakly-correlated, 2-dimensional model, moderately misspecified models: IQR = [10.580, 22.366]; median = 16.048.

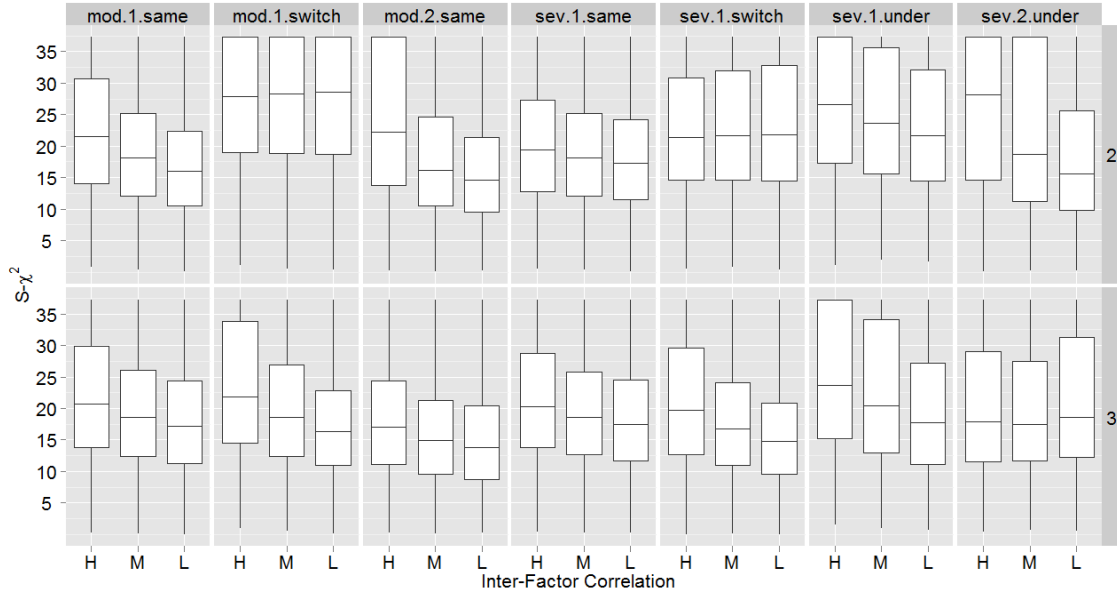


Figure 5.8. Box-and-Whiskers Plots for the $S\text{-}\chi^2$ Under Model Misspecification.

Presented according to conditions associated with sensitivity: inter-factor correlation, number of dimensions (rows), and type of misspecification (columns). Type of misspecification presented as degree of misspecification (Moderate, Severe), item multidimensionality (Between = 1, Within = 2), and item misspecification (Same = Correct, Switch = Alternate-factoring, Under = Underfactoring).

5.3.1.2 Distributional Characteristics of the Modification Index

The Modification Index (MI) indicates the approximate decrease in model-fit χ^2 if the current parameter were freely estimated. For the purpose of identifying model misspecification, MI values indicate Q-matrix elements that would improve model fit if the item were associated with the latent factor. MI values are, therefore, separately estimated for each of the 2 or 3 latent factors (i.e., MI1, MI2, and MI3). Since MI3 can only be calculated for models containing 3 latent factors, number of dimensions is excluded from the factorial ANOVA when calculating sensitivity. Otherwise, the patterns of sensitivity are seen to be similar across Modification Indices; given the similarity of the patterns, subsequent discussion is limited to MI1 in an effort to the complexity of analysis and interpretation. Unique variation is seen to account for the majority of variance ($\eta^2 = 68.911\%$) followed by sample size ($\eta^2 = 4.332\%$), number of dimensions

($\eta^2 = 3.035\%$), and inter-factor correlation ($\eta^2 = 3.536\%$). This pattern is the same as the pattern of sensitivity demonstrated under true model estimation. The distribution of the Modification Index 1 is presented in Figure 5.9 in a 90%-winsorized box-and-whiskers plot according to sample size, number of dimensions, and inter-factor correlation. Values of MI1 are seen to increase with sample size and decrease with number of dimensions and strength of inter-factor correlations. The largest values of MI1, indicating misspecification, are demonstrated when weakly-correlated 2-dimensional models are estimated with 1000 examinees: IQR = [12.033, 42.221]; median = 21.580. The smallest values of MI1 are demonstrated when highly-correlated 3-dimensional models are estimated with 250 examinees: IQR = [0.106, 1.369], median = 1.0861.

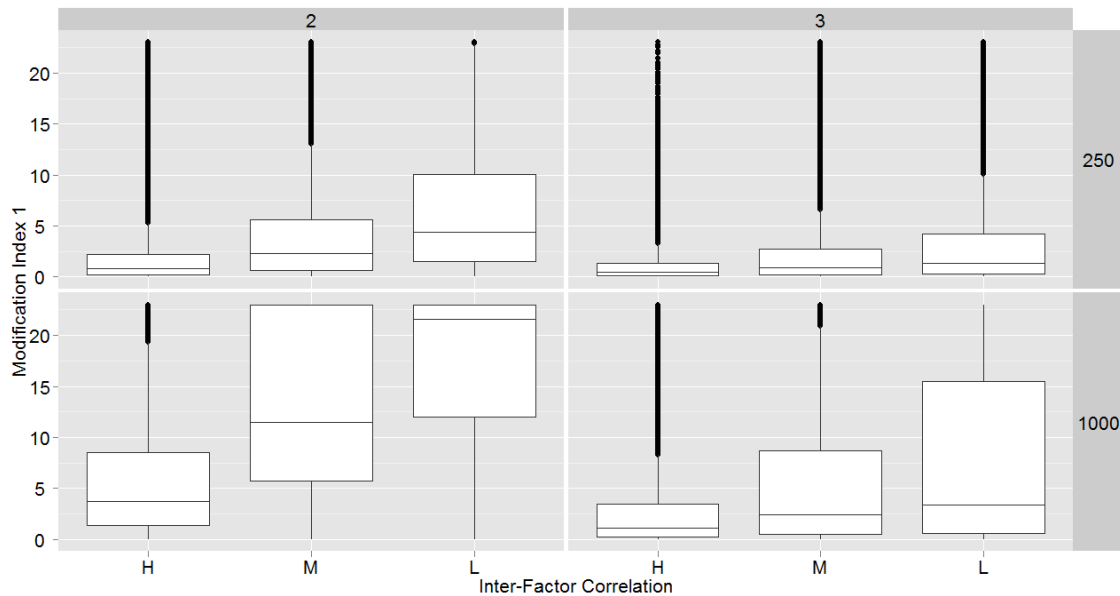


Figure 5.9. Box-and-Whiskers Plots for Modification Index 1.

Presented according to conditions associated with sensitivity: inter-factor correlation, sample size (rows), and number of dimensions (columns).

5.3.1.3 Distributional Characteristics of the Wald Test

Like the Modification Indices, the Wald Test demonstrates similar patterns of sensitivity across latent factors, therefore, only Wald Test 1 will be discussed. Before interpreting the Wald Test values it is important to recall that this item-fit statistic is used to test significance of specific factor loadings; smaller values suggest misspecification indicating that the estimated factor loading, or Q-matrix entry, is non-significant. Keeping all this in mind, the Wald Test is seen to demonstrate sensitivities strikingly similar to the pattern and magnitude seen under true, correct model specification; a large portion of total variance in this fit index is attributable to the compound factor of misspecification type ($\eta^2 = 40.534\%$), which includes item multidimensionality – the largest source of variance in the Wald Test values under true model estimation; lesser percentages of variance are attributed to sample size ($\eta^2 = 18.893\%$) and item type ($\eta^2 = 10.091\%$). Depicted in Figure 5.10, values of Wald Test 1 appear to generally decrease with item discrimination, item difficulty, sample size, type of misspecification (alternate-factoring and underfactoring). Values of the Wald Test are smallest, suggesting misspecification, when high discrimination / high difficulty items are modeled as within-item multidimensional and when they estimated as underfactored within severely misspecified models and small sample sizes: IQR = [-0.569, 1.896]; median = 0.835.

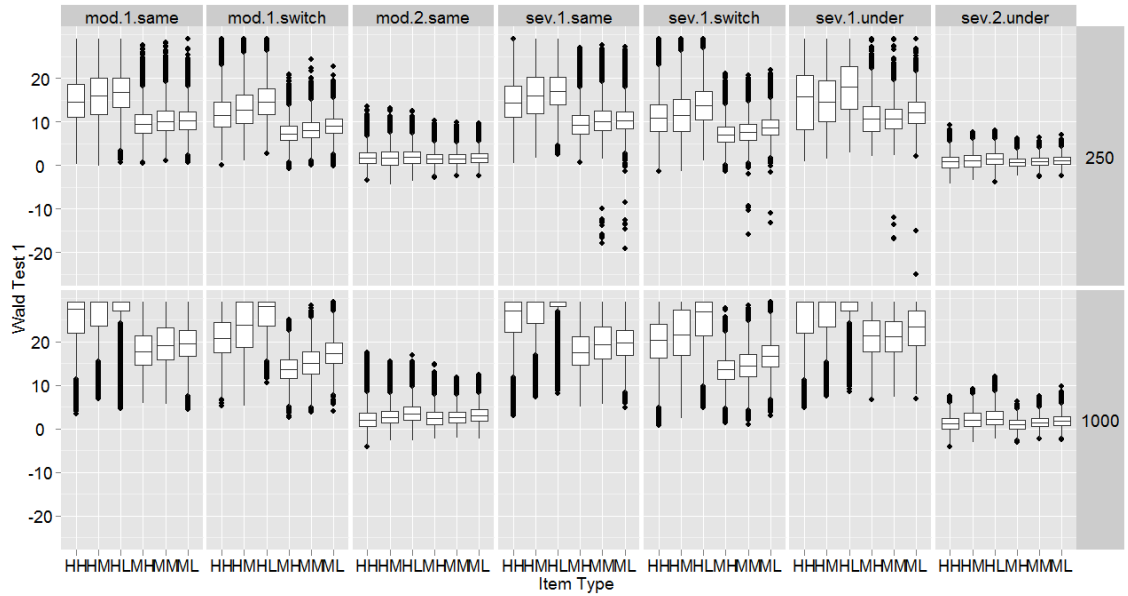


Figure 5.10. Box-and-Whiskers Plots for Wald Test 1.

Presented according to conditions associated with sensitivity: item type, sample size (rows), and type of misspecification (columns).

5.3.2 Power of Item-fit Indices

In this dissertation, items were either correctly estimated or misspecified as being associated with an alternate factor or underfactoring via the deletion of a Q-matrix entry, within each of the model misspecification conditions. This allows for the calculation of power as the average rejection rate for misspecified items calculated by aggregating over item type and across replications. Rejection results from correct identification of an item as misfitting via the application of design-appropriate empirical cut points, which were calculated as the 95th percentile for each of the $S-\chi^2$, Modification Indices, and Wald Tests according to each cell in the simulation design. These proportions are then computed for each cell in the simulation design, aggregating over replications. The sensitivity of each item-fit index's ability to correctly reject misspecified items is presented in Table 5.9 as the percentage of variance attributable to the simulation

conditions and interactions resulting from factorial ANOVAs conducted for each fit index.

Table 5.9
Selected Percentages of Variance for Power of Item-Fit Statistics

Source	S- χ^2	Modification Index			Wald Test		
		1	2	3	1	2	3
Number of Dimensions (1)	16.218	4.642	12.398		13.567		
Test Length (2)	6.135	1.360	0.006	2.862	0.050	0.013	1.559
Sample Size (3)	21.840	20.293	16.152	15.325	0.770	0.047	0.354
Inter-factor Correlation (5)	1.711	12.512	8.067	8.397	19.237	25.335	12.921
Item Type (6)	9.760	4.458	3.033	3.850	11.262	9.123	29.265
Misspecification Type (7)	8.901	3.144	12.727	15.231	11.551	17.097	19.681
1*2	0.861	3.758	0.211		2.221		
1*3	0.001	1.813	2.907		0.030		
1*5	2.175	0.983	1.238		0.697		
1*6	3.015	0.378	1.097		8.083		
1*7	5.383	0.179	1.720		1.134		
2*3	0.876	1.607	0.151	0.004	0.138	0.000	0.334
2*7	1.952	0.997	5.163	19.257	0.976	0.119	0.293
3*5	0.348	0.892	0.305	1.652	1.156	1.392	0.143
3*7	4.840	0.066	0.851	2.666	1.379	0.399	1.186
5*6	0.688	0.092	0.098	0.251	3.041	9.008	14.473
5*7	0.743	0.123	0.818	3.165	3.339	16.592	0.655
6*7	1.402	0.079	0.165	0.779	5.825	6.201	12.293
1*2*6	0.109	0.128	0.292		2.711		
1*2*7	0.353	11.613	11.619		0.194		
1*3*7	1.829	0.639	1.323		0.141		
1*6*7	0.820	0.171	0.434		1.216		
2*6*7	0.657	0.189	0.337	2.179	1.336	0.313	0.776
3*6*7	1.383	0.064	0.318	0.445	0.646	0.340	0.157
5*6*7	0.144	0.109	0.164	0.281	1.783	5.890	1.889
Residuals	1.866	22.545	5.968	8.685	1.042	2.050	0.727

* Cells highlighted in dark grey indicate top-three sources of variance; cells highlighted in light grey indicate top main effects suggested by top-three interactions.

5.3.2.1 Power of $S\text{-}\chi^2$

The largest percentage of variance in proportion of misfitting items correctly rejected by the $S\text{-}\chi^2$ can be attributed to sample size ($\eta^2 = 21.840\%$), next is the number of dimensions or latent factors ($\eta^2 = 16.218\%$), and lastly item type ($\eta^2 = 9.760\%$). Power of the $S\text{-}\chi^2$ item-fit index to detect misspecified items is presented in Figure 5.11 and summarized according to those simulation conditions for which it demonstrated sensitivity. Power is seen to increase with sample size and item discrimination but decrease with number of dimensions and item difficulty. Power is highest for 2-dimensional models with highly-discriminating / low-difficulty items estimated on large sample sizes ($n = 1000$), $\text{IQR} = [0.576, 0.890]$ and median = 0.714, while power is lowest for 3-dimensional models and sample sizes of $n = 250$ where the $S\text{-}\chi^2$ is shown to rarely detect item misspecification, with power rates approaching zero.

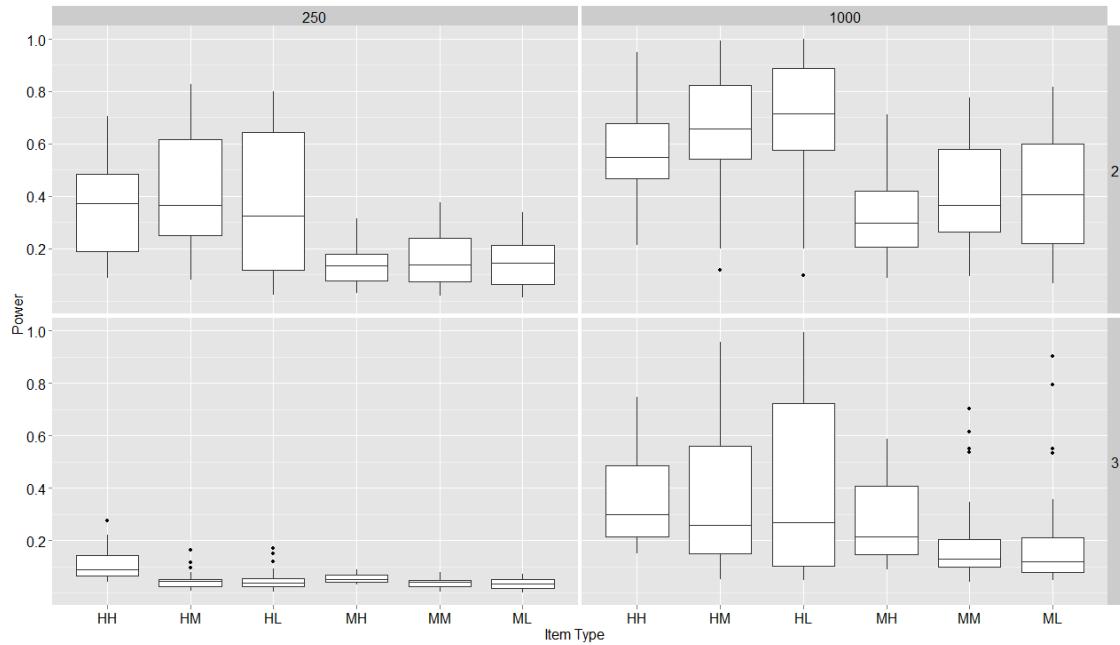


Figure 5.11. Box-and-Whiskers Plots for Power of the $S\text{-}\chi^2$.

Presented according to conditions associated with sensitivity: item type, number of dimensions (rows), and sample size (columns).

5.3.2.2 Power of the Modification Index

Power of the Modification Index is shown to be sensitive to a variety of simulations conditions and the interactions thereof as well as differing depending on the latent factor being considered. It is also important to consider that the Modification Indices are estimated as a result of specific Q-matrix properties. Modification Index values are estimated for null (“0”) entries in the estimated Q-matrix; MI1, therefore, directly results from actual null Q-matrix entries as well as alternate- and underfactoring of the specific Q-matrix element; MI2 is similar to MI1 except under 3-dimensional models where it only directly results from null entries and underfactoring of the Q-matrix; MI3 results from all null Q-matrix entries but only present when 3-dimensional models are estimated.

MI1 (Figure 5.12) is shown to be sensitive to sample size ($\eta^2 = 20.293\%$), inter-factor correlation ($\eta^2 = 12.512\%$), and the second-order interaction of number of dimensions, test length, and type of misspecification ($\eta^2 = 11.613\%$), plus a variety of other conditions to a lesser degree. MI2 (Figure 5.13) demonstrates sensitivity to sample size ($\eta^2 = 16.152\%$), type of misspecification ($\eta^2 = 12.727\%$), and number of dimensions ($\eta^2 = 12.398\%$). Lastly, MI3 (Figure 5.14) demonstrates sensitivity to the interaction of test length and type of misspecification ($\eta^2 = 19.257\%$), sample size ($\eta^2 = 15.325\%$), and the main effect of misspecification type ($\eta^2 = 15.231\%$). Though the three Modification Indices demonstrate different sensitivities and power rates there are overall patterns that can be observed. Power is seen to increase with sample size, degree of model misspecification, and item misspecification – larger values for underfactoring than alternate-factoring. MI1 demonstrates the highest consistent power for weakly-correlated 2-dimensional models estimated with large sample sizes: IQR = [0.724, 0.978] and median = 0.869. Alternately, power decreases with number of dimensions estimated, such that alternate-factoring under 3-dimensional models with small sample sizes results in power rates approaching zero.

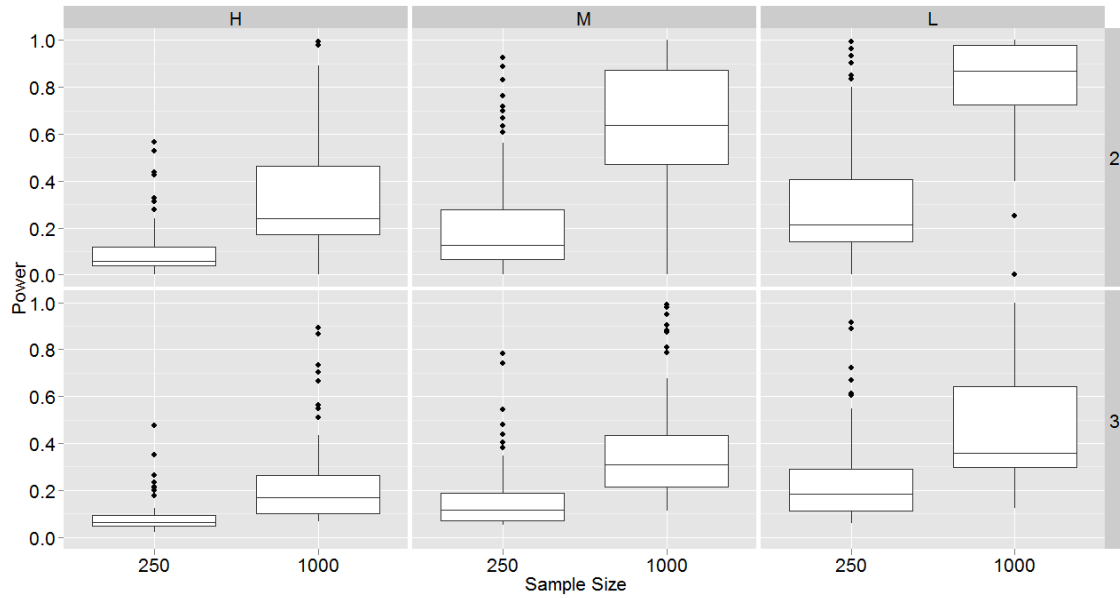


Figure 5.12. Box-and-Whiskers Plots for Power of the Modification Index 1.

Presented according to conditions associated with sensitivity: sample size, number of dimensions (rows), and inter-factor correlation (columns).

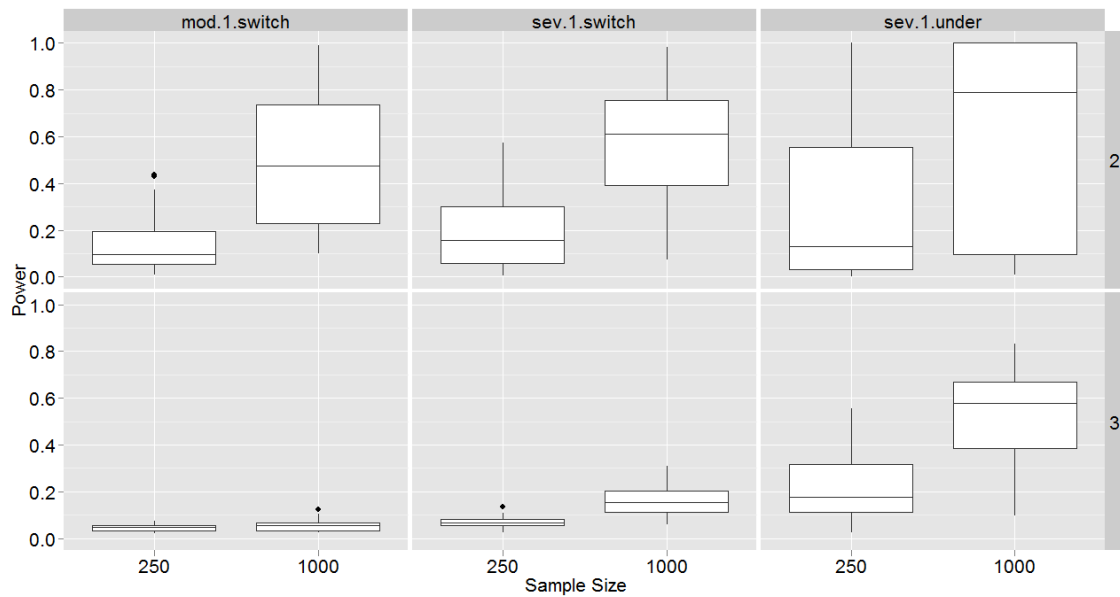


Figure 5.13. Box-and-Whiskers Plots for Power of the Modification Index 2.

Presented according to conditions associated with sensitivity: sample size, number of dimensions (rows), and type of misspecification (columns).

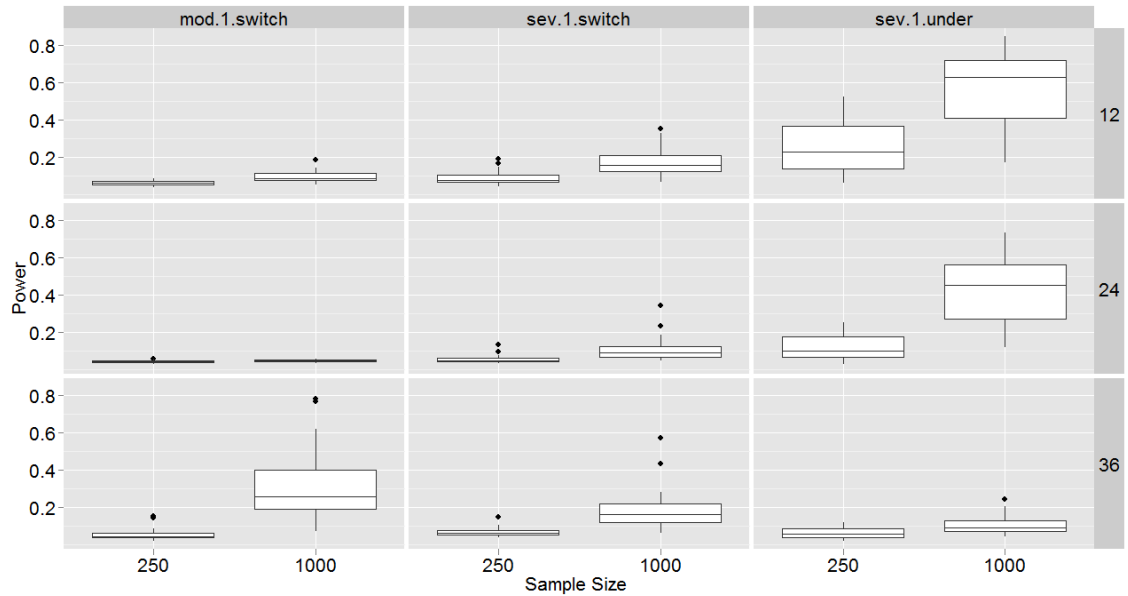


Figure 5.14. Box-and-Whiskers Plots for Power of the Modification Index 3.

Presented according to conditions associated with sensitivity: sample size, test length (rows), and type of misspecification (columns).

5.3.2.3 Power of the Wald Test

When using the empirically-derived cut points, the power rates for the Wald Test are generally shown to be low across conditions for all three indices. The patterns of sensitivity across Wald Test 1, Wald Test 2, and Wald Test 3 demonstrate notable similarities, with the simulation conditions accounting for approximately 90% of the variance in each statistic. For Wald Test 1, the largest percentages of variance is attributed to inter-factor correlation ($\eta^2 = 19.237\%$), the number of dimensions ($\eta^2 = 13.567\%$), and the type of misspecification ($\eta^2 = 11.551\%$). Wald Test 2 also demonstrates great sensitivity to inter-factor correlation ($\eta^2 = 25.335\%$), the main effect of type of misspecification ($\eta^2 = 17.097\%$) and the interaction of inter-factor correlation with type of misspecification ($\eta^2 = 16.592\%$), as well as demonstrating sensitivity to item type ($\eta^2 = 9.123\%$). There is no effect of number of dimensions for Wald Test 2 since

misspecified items are only associated with latent factor 1 or 3. Finally, Wald Test 3 is sensitive to inter-factor correlation ($\eta^2 = 12.921\%$, via interaction with item type), type of misspecification ($\eta^2 = 19.681\%$), and item type ($\eta^2 = 29.265\%$).

Across the Wald Test item-fit indices, power rates are seen to increase with the number of dimensions and the severity of model misspecification while decreasing with inter-factor correlation, item discrimination, and item difficulty. Power rates are highest for Wald Test 1 when items are severely misspecified according alternate-factoring within a weakly-correlated 3-dimensional model: IQR = [0.448, 0.784], median = 0.593. The lowest power rates are observed when high-difficulty / high-discrimination items are underfactored within severely-misspecified highly-correlated 2-dimensional models.

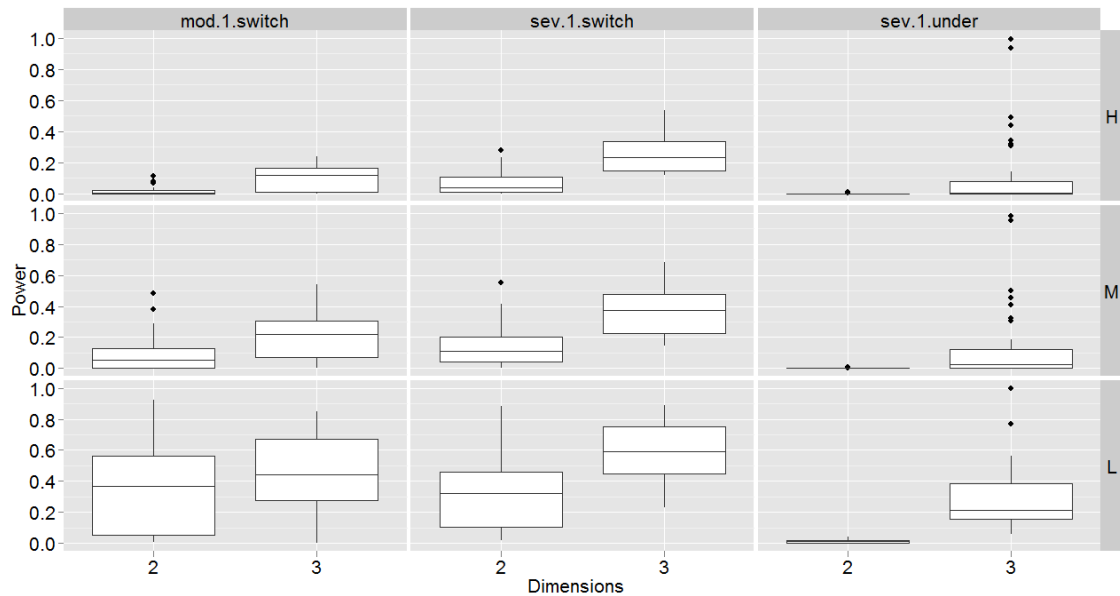


Figure 5.15. Box-and-Whiskers Plots for Power of the Wald Test 1.

Presented according to conditions associated with sensitivity: number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

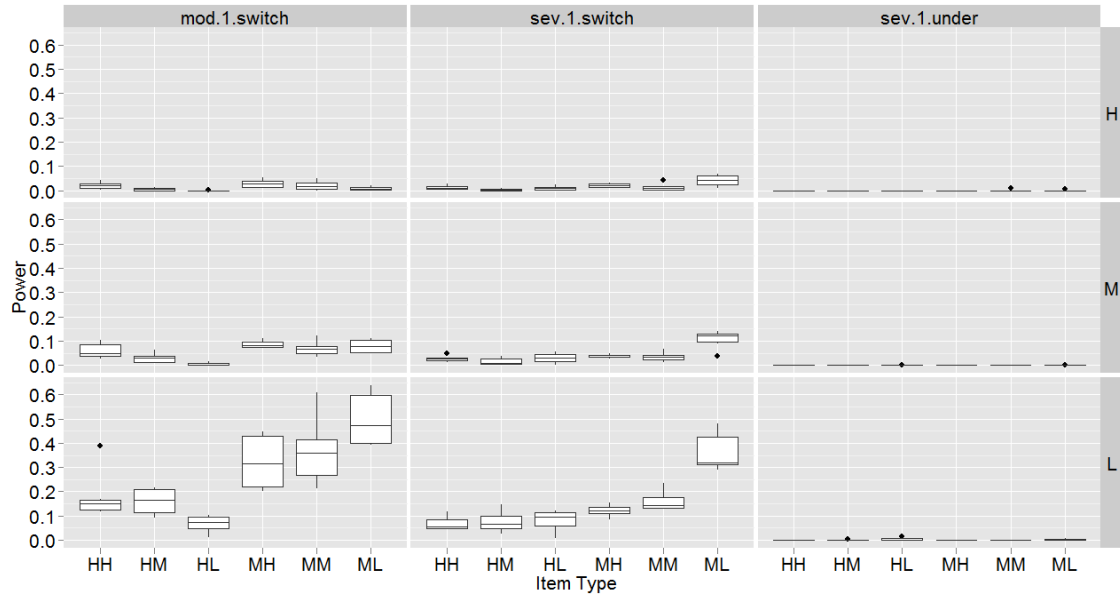


Figure 5.16. Box-and-Whiskers Plots for Power of the Wald Test 2.

Presented according to conditions associated with sensitivity: item type, inter-factor correlation (rows), and type of misspecification (columns).

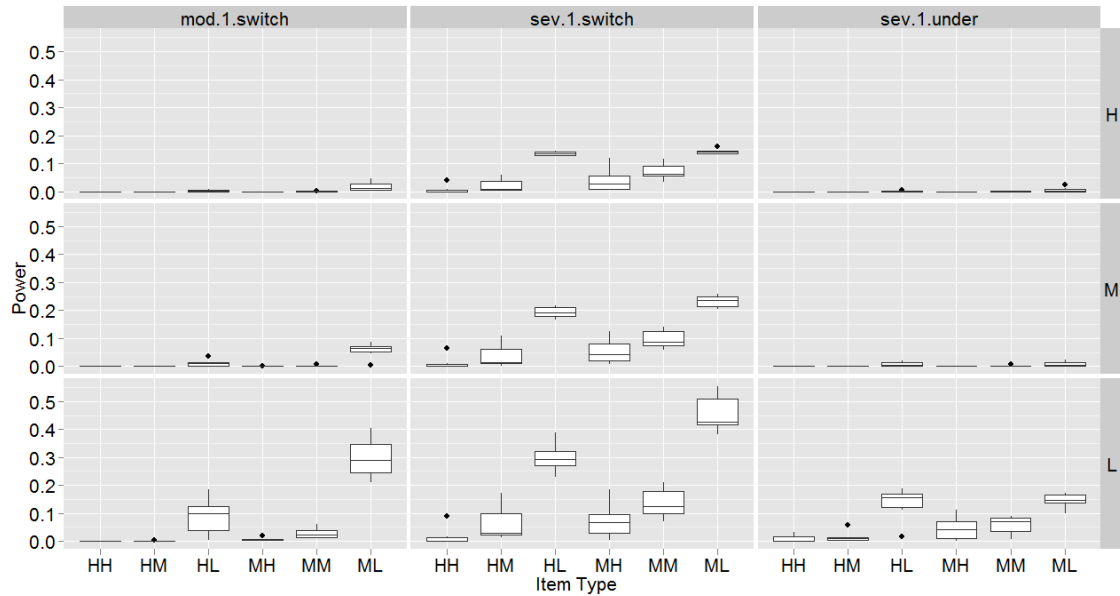


Figure 5.17. Box-and-Whiskers Plots for Power of the Wald Test 3.

Presented according to conditions associated with sensitivity: item type, inter-factor correlation (rows), and type of misspecification (columns).

5.3.3 Summary for Item-Fit Indices

Overall, the item-fit indices often demonstrated the ability to correctly reject misspecified items when there were 2 weakly correlated latent factors, large sample sizes, and items of low difficulty. The item-fit indices demonstrated poor ability to detect misspecified items under strongly-correlated 3-dimensional models which were estimated on small sample sizes. The $S-\chi^2$, Modification Indices, and Wald Test statistics each demonstrated variable power rates with respect to simulation conditions, described earlier and summarized below. Considering the estimation issues described at the beginning of this chapter, the following results often correspond with the conditions resulting in estimation difficulties, however, the two are not completely aligned, indicating that estimation issues do not entirely account for the observed effects.

Power of the $S-\chi^2$ is seen to be highest 2-dimensional models estimated on large sample sizes with items of high discrimination / low difficulty. These rates range typically between 0.3 and 0.9 and are comparable to those of Li and Rupp (2011) who found power to be 0.4 and 0.8 for moderate and high inter-factor correlations when data generated according to a 2-dimensional 2PL-MIRT model was estimated according to a unidimensional model – that is, subject to underfactoring. Also, Zhang and Stone (2008) found the power to detect misspecification using the $S-\chi^2$ to range 0.7 to 0.93 for items estimated according to a 2PL-MIRT model and misspecified as violating the assumption of monotonicity.

Similar to the $S-\chi^2$, Modification Indices show power rates that are highest under weakly-correlated 2-dimensional models and when test are short in length. Under these conditions, power rates vary between approximately 0.4 and 0.8, while the extreme

opposite conditions demonstrate power rates that approach zero. Previous research has shown model revision and recovery of the correct population model via Modification Indices to be moderately successful under large sample sizes when misspecification is moderate (Kaplan, 1990; MacCallum, 1986). These findings correspond with the modest power rates especially under large sample sizes found in this dissertation.

Similar to the previous item-fit indices, the Wald Test statistics demonstrate the highest power rates when inter-factor correlation is low; unlike the previous item fit indices, however, power rates for the Wald Test increase with the number of latent factors. Additionally, power is seen to increase with severe alternate-factorings while decreasing with item discrimination and difficulty. Chou and Bentler (2002) found that the Wald Test correctly indicated misspecified parameters in 88 out of 100 instances when a saturated 5-dimensional CFA model was estimated and the Wald Test was examined to suggest parameter deletion in an attempt to recover the true population model. These results are aligned with the expectations for the Wald Test presented in this dissertation.

Considering the results of presented for these three item-fit indices it is important to note that the $S-\chi^2$ and the Modification Indices demonstrated sensitivities to observable design characteristics such as sample size and number of dimensions – where power increases with the former and decreases with the latter. The Wald Test, however, is typically sensitive to those unobserved characteristics that would only be discovered upon model estimation. Lastly, we see that the Modification Indices are able to detect underfactoring at high rates and severe alternate-factorings results in increased power rates

for both the Modification Indices and the Wald Test; the $S-\chi^2$, however, was less sensitive to degree and type of misspecification than the other simulation conditions described.

5.4.Synthesis of Model- and Item-Fit Performance Under Model Misspecification

In the evaluation of model and item fit under conditions of potential misspecification, it is important to understand the sensitivity of the fit indices to the experimental or simulation conditions currently employed. Appropriate consideration of the effects of model and test characteristics on the selected model- and item-fit indices will allow modelers – practitioners and researchers, alike – to make appropriate decisions when considering model validity and revision. Adequate power to correctly detect model misspecification is generally demonstrated by the χ^2/df ratio, RMSEA, and GDDM model-fit indices, with certain exceptions such as when sample sizes are small and short tests are employed. Power to detect item misspecification by the $S-\chi^2$, Modification Index, and Wald Test item-fit indices, however, is quite variable demonstrating power rates that are often low. Since model- and item-fit indices are both typically presented in model estimation output – for example, Mplus version 6.11 (Muthén & Muthén, 1998-2010) can output necessary information for the χ^2/df ratio, RMSEA, Modification Index, and Wald Test– the power of the two types of indices are next considered in conjunction for the purpose of providing additional information and guidance regarding identification of item and model (i.e., Q-matrix) misspecification. The power of each item-fit index to correctly identify misspecified items was calculated for each model-fit index separately, according to whether the model-fit index correctly identified the misspecified model.

5.4.2 Misspecification Correctly Detected by Model Fit Indices

Seen in Figure 5.18, correct model rejection according to the χ^2/df results in modest increases in rejection of misspecified items by the $S\text{-}\chi^2$, though the power rates are still modest overall and poor for 3-dimensional models. After allowing for a modest increase, the pattern and magnitude of power to detect misspecified items using the $S\text{-}\chi^2$ when the models were identified as misspecified is remarkably similar to when model fit was not considered. For example, the IQR for power rates under 2-dimensional models estimated for 1000 examinees and highly-discriminating / low difficulty items was approximately 0.6 to 0.8 overall but increases to 0.6 to 1.0 when the χ^2/df first identifies the model as misspecified. A similar effect is seen for the Modification Index (MI1 is presented for ease of interpretation); the patterns of power rates are similar to the overall pattern but the initial identification of the misspecified model results in increased power rates for the item-fit statistic overall. Power for the Modification Index is highest for 2-dimensional models estimated as weakly-correlated with large samples sizes, with an IQR ranging approximately 0.7 to 1.0; after successful identification by χ^2/df , the IQR increases to approximately 0.8 to 1.0, with a median of 1.0. The Wald Test, also limiting presentation to latent factor 1, appears to be least affected by initial identification of model misspecification as power rates are seen to differ little from the overall power rates. For weakly-correlated, severely misspecified, 3-dimensional models the IQR for underfactored items was approximately 0.4 to 0.8 (demonstrating the highest power overall) which increases to 0.3 to 0.85 subsequent to identification by the χ^2/df model fit index.

Figure 5.19 shows the power rates for the $S\text{-}\chi^2$, MI1, and Wald Test 1 subsequent to correct identification of model misspecification by the RMSEA. The previously established similarities in performance between the χ^2/df and RMSEA model fit indices again provide nearly identical results; generally, the magnitude of the power rates is increased but the overall pattern of power is maintained.

Lastly, Figure 5.20 presents the power rates for the item-fit indices under correct identification of model misspecification by the GDDM model-fit index. As is apparent, these results are similar to those previously presented and require no further discussion.

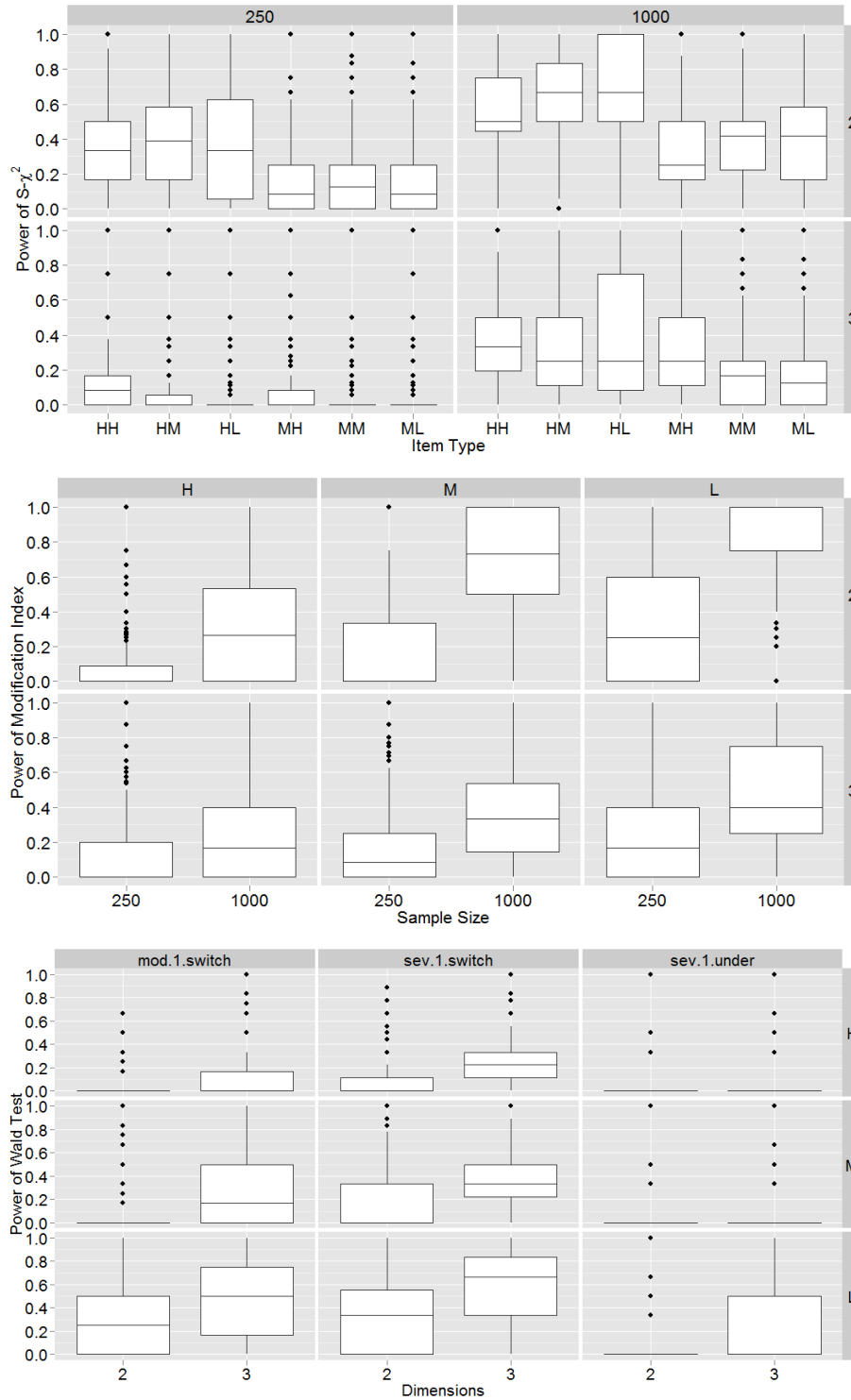


Figure 5.18. Power of item fit indices when χ^2/df ratio correctly indicates model misfit.

$S-\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

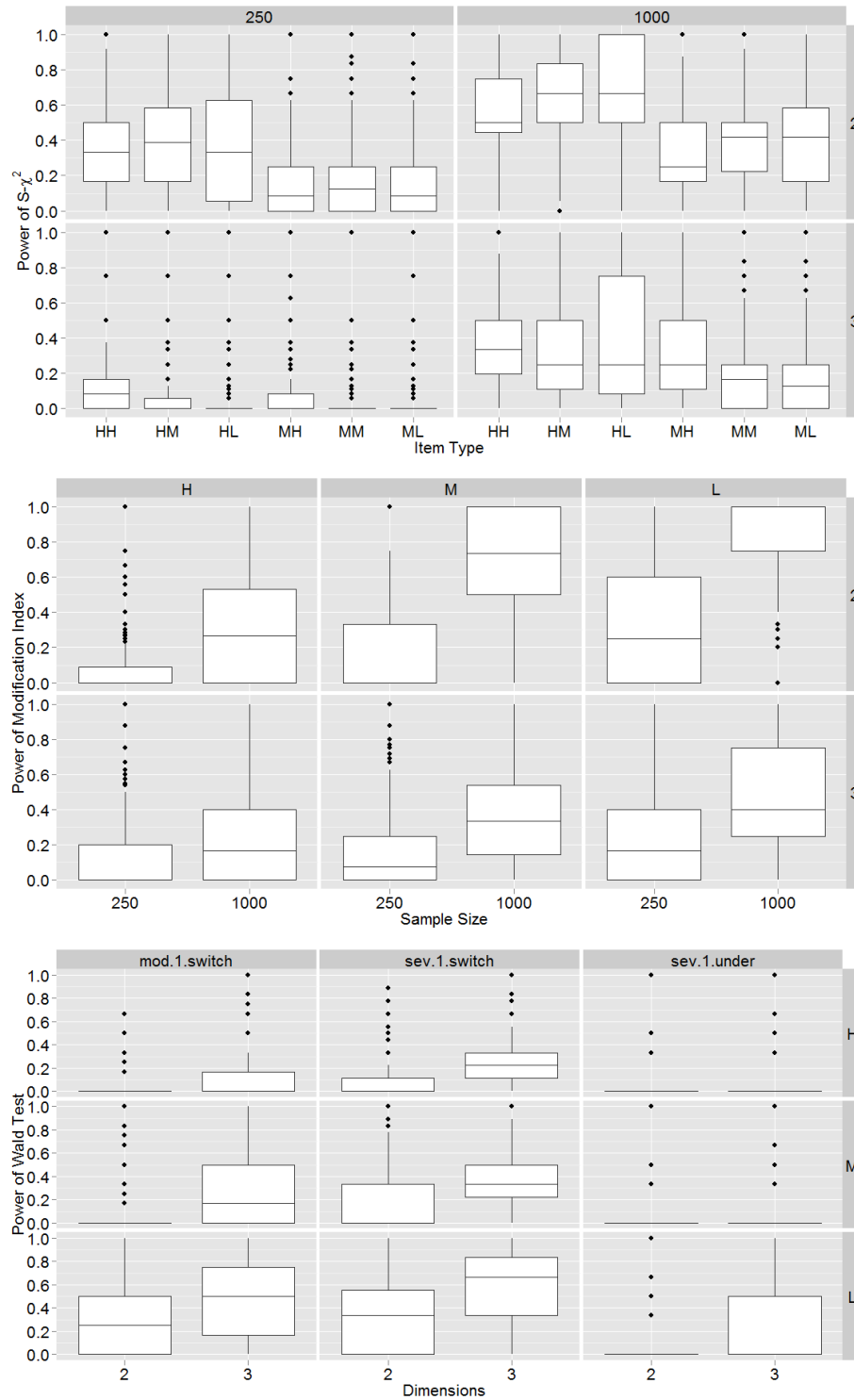


Figure 5.19. Power of item fit indices when RMSEA correctly indicates model misfit.

$S-\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

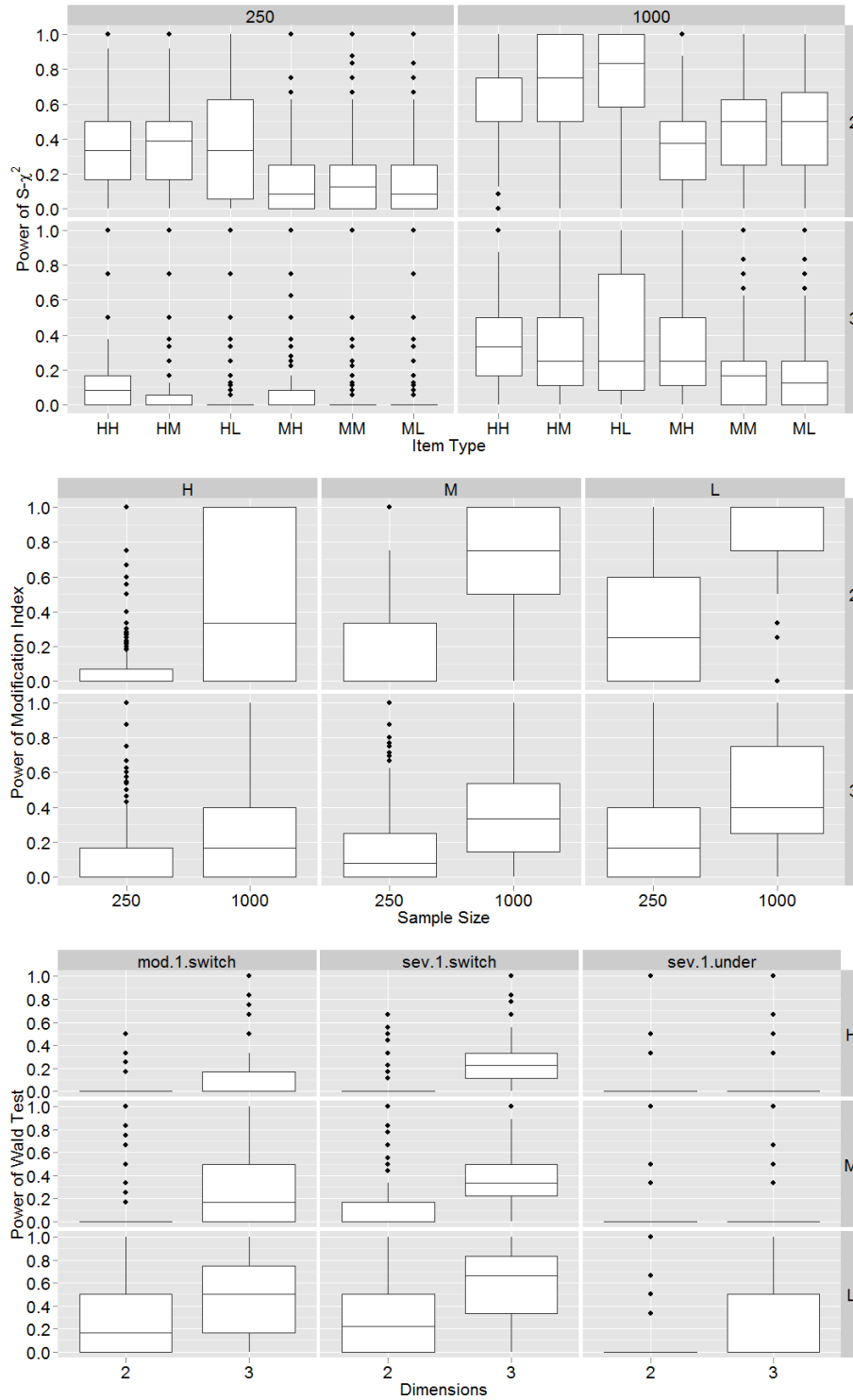


Figure 5.20. Power of item fit indices when GDDM correctly indicates model misfit.

$S-\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

5.4.3 Misspecification Not Detected by Model Fit Indices

When the χ^2/df , RMSEA, or GDDM model fit indices unsuccessfully reject a misspecified model, Figure 5.21 through Figure 5.23 shows that the item-fit indices subsequently reject misspecified items at rates generally lower than when the model-fit indices successfully rejected misspecified models, though evidencing the same patterns. There are a few instances where distributions of power rates are missing from the figures, indicating that all models within that combination of simulation conditions were correctly rejected by the model-fit index. For example, all weak- and moderate-correlated models with sample sizes of 1000 were correctly identified as misspecified by χ^2/df .

There are also some instances where the performance of the item-fit statistics deviates from the description above. For items of moderate discrimination and low-to-moderate difficulty estimated under 3-dimensional models, the inter-quartile ranges and median power rates for the $S\text{-}\chi^2$ increase when the misspecified model is unidentified by the χ^2/df (approximate median for identified = 0.15; approximate median for unidentified = 0.25) and the RMSEA (identified = 0.15; unidentified = 0.25), but not for the GDDM.

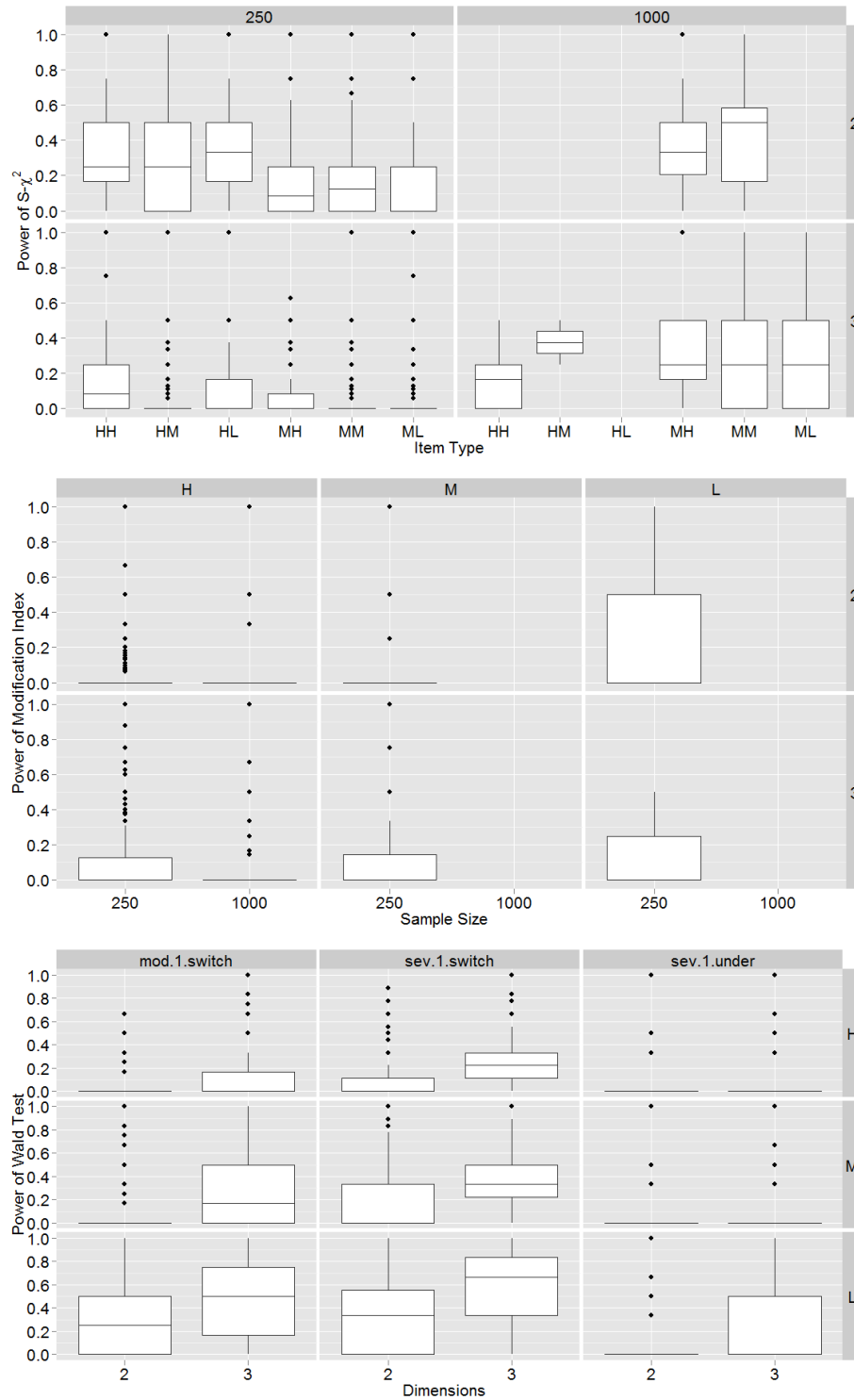


Figure 5.21. Power of item fit indices when χ^2/df ratio fails to indicate model misfit.

$S\text{-}\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

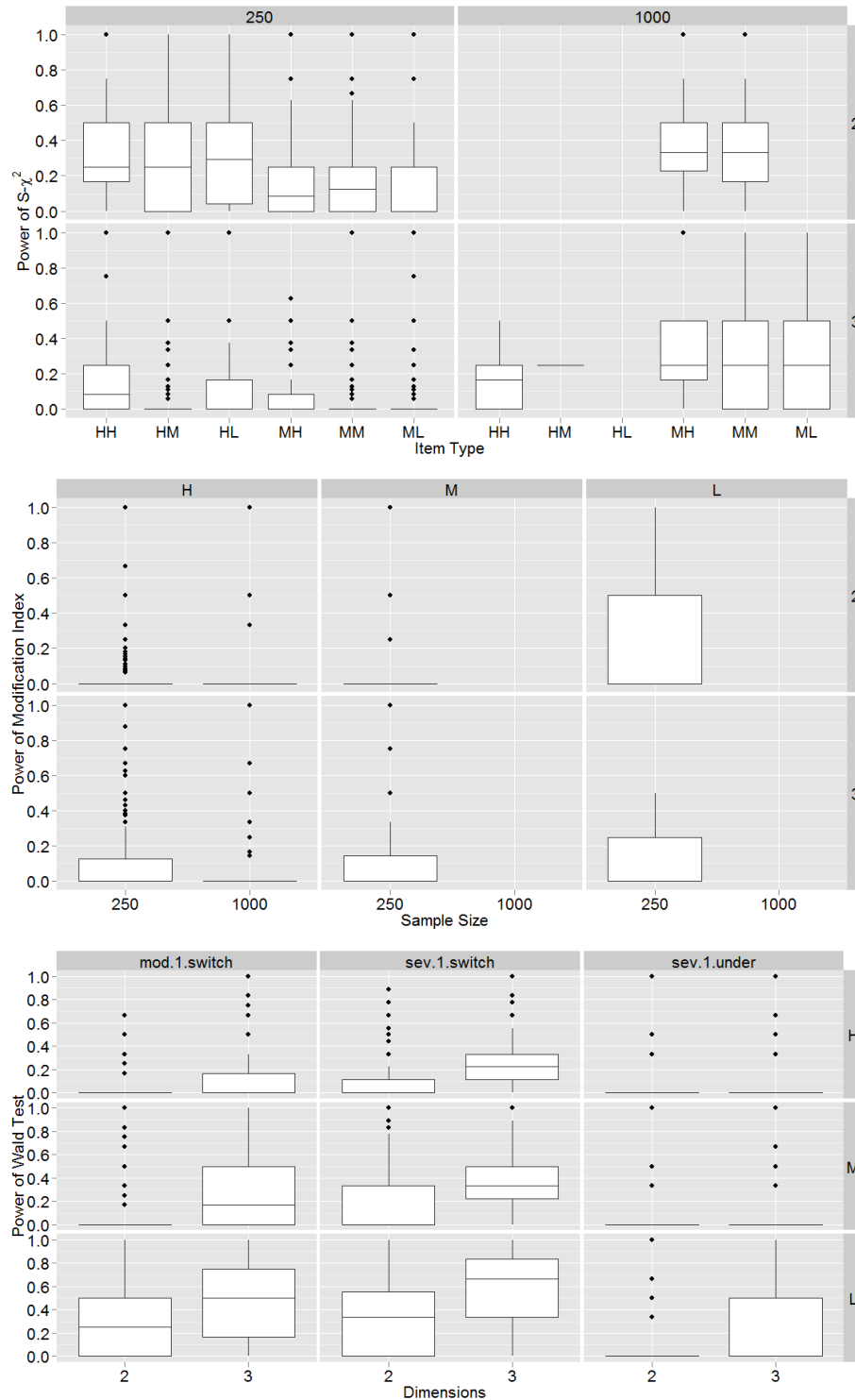


Figure 5.22. Power of item fit indices when RMSEA fails to indicate model misfit.

$S-\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

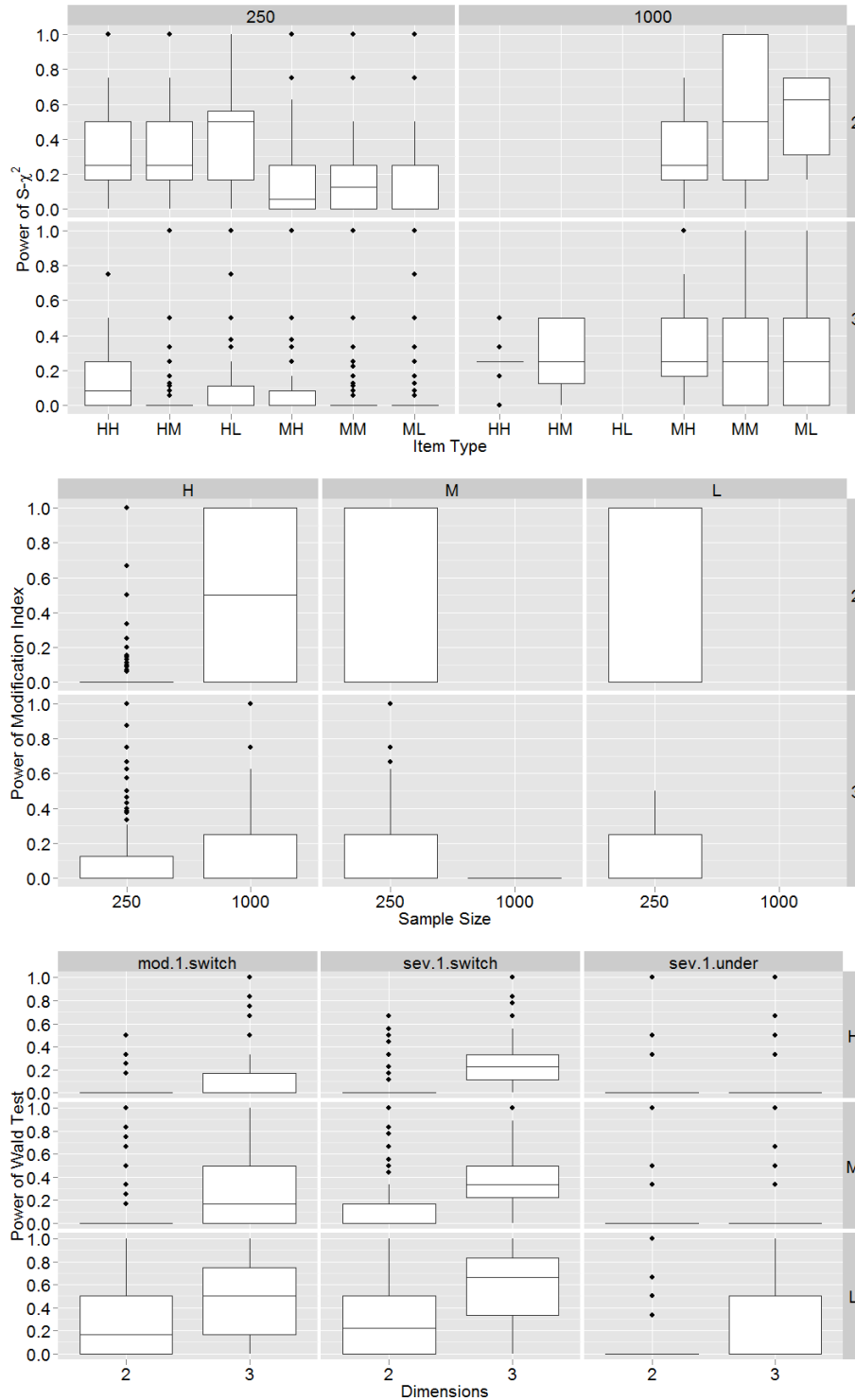


Figure 5.23. Power of item fit indices when GDDM fails to indicate model misfit, for between-item multidimensional items.

$S\text{-}\chi^2$ (top) is presented according to item type, number of dimensions (rows), and sample size (columns). Modification Index 1 (middle) is presented according to sample size, number of dimensions (rows), and inter-factor correlation (columns). Wald Test 1 (bottom) is presented according to number of dimensions, inter-factor correlation (rows), and type of misspecification (columns).

5.5.Summary

This section presented a unified approach to the evaluation of model misspecification, considering model- and item-fit results simultaneously. As these statistics are often presented or otherwise available together during model estimation, joint evaluation provides richer data with which to judge the Q-matrix specifying the measurement model or pattern of factor loadings. When the model-fit indices correctly rejected the models, the ability of the item-fit indices to reject or identify misspecified items generally increased. Power rates for the model-fit indices uniformly increased with larger samples, longer test lengths, and weaker inter-factor correlations. Alternately, failure to correctly reject misspecified models resulted in slightly decreased power rates compared to those demonstrated when model evaluation was not initially considered, with the exception of the $S-\chi^2$ under 3-dimensional models which demonstrated slight improvements. These results suggest that the information provided by the $S-\chi^2$, Modification Index, and Wald Test item-fit statistics is consistent, regardless of whether the model-fit statistic was able to detect misspecification. A further implication of this behavior is that item-fit indices can validly be used during model criticism and evaluation procedures, even when the overall model was judged to fit the data.

All of the item-fit indices demonstrated power ranging from poor to strong, depending on the simulation conditions considered. Having initially identified a misspecified model as such, the $S-\chi^2$ is able to detect misspecified items with a power of greater than 0.5 when sample sizes are large, the model is estimated as 2-dimensional, and items are highly-discriminating. The Modification Index is able to detect misspecified items with a power greater than 0.5 when sample sizes are large and latent

factors are weakly correlated or moderately correlated, though only for 2-dimensional models. The Wald Test, however, is only able to detect misspecified items with a power of 0.5 or greater when a weakly-correlated 3-dimensional is severely misspecified and the items have been subject to alternate-factoring. When the overall model has not been identified as misspecified, the utility of the $S-\chi^2$ is similar to that described above, though power is slightly lessened overall; the utility of the Modification Index is limited to 2-dimensional models, only; and Wald Test continues to demonstrate lower power rates. The poor performance of the Wald Test can be attributed to the fact that the estimated values are typically large and range widely; calculated as the ratio of the factor loading to the standard error of the estimate, it may be surmised that the Wald Test would be more informative in detecting misspecification for items with lower factor loadings or item discrimination values than those specified in this dissertation.

Chapter 6

Real Data Analysis

6.1. Introduction

The final research question posed in this dissertation is in regards to the application of findings from the simulation studies to real data:

How can results from the simulation studies inform model criticism and model revision for real data analysis contexts when Q-matrices are potentially misspecified?

The Q-matrices employed in estimating these models are constructed according to (1) the results of exploratory factor analyses (EFA) and (2) the assignment of test items to levels of the revised *Bloom's Taxonomy for Educational Objectives* (Anderson & Krathwohl, 2001; Bloom et al., 1956).

Two-parameter normal-ogive (2-PNO) multidimensional item response theory (MIRT) models are then estimated using item response data from a grade 6 mathematics achievement assessment administered in a large Midwestern state. These models are specified according to the aforementioned Q-matrices; test and sample characteristics resulting from each of the estimated models are then examined for correspondence to conditions employed in the simulation study portion of this dissertation. Design-appropriate empirical cut points resulting from simulation conditions best approximating the real data analysis conditions are then applied to the model- and item-fit indices for the purpose of adjudicating fit. Further, lessons learned about the behavior and power of the fit indices under various simulated test design and model estimation conditions are

incorporated in evaluating overall model fit and suggesting Q-matrix revisions. A single iteration of model revision is presented for illustrative purposes.

6.2.Methods

The full data set represents the population of students in the state and is comprised of 12,861 students' responses to 39 multiple-choice and constructed response items; for this analysis, a random sample of 1000 examinee responses to the 32 multiple-choice items only are included, thus focusing on dichotomously-scored responses and approximating a sample size and test length condition in the simulation study.

Model estimation according to EFA-derived Q-matrices first required that the number of latent factors be determined. To account for potential sampling bias, the number of factors was determined using Horn's *Parallel Analysis* (1956) method, implemented in R as `psych::fa.parallel.poly` (Revelle, 2011), for 250 random samples of $n = 1000$ drawn from the population data. The number of factors extracted ranged from 2 to 13 with mean, median, and mode all suggesting a six-dimensional model which is larger than any of the Q-matrices used in the simulation study. The data set yielding this six-dimensional solution was retained and employed in all subsequent analyses.

To facilitate the use of the empirical cut points determined in the simulation study a two-dimensional EFA solution is also considered in this analysis. The two- and six-dimensional Q-matrices were then constructed from the EFA results by estimating oblique two- and six-dimensional factor solutions and defining $q_{jk} = 1$ as the one or two largest positive factors loadings across dimensions. This method of Q-matrix construction ensures that item-dimensionality is similar to that of the simulation study as well as

capturing positive relationships between the observed and latent variables, as would be expected for MDISC values. The resulting two- and six-dimensional Q-matrices follow complex-structure and show high proportions of items which are within-item multidimensional; 21 of the 32 items in the two-dimensional solution are within-item multidimensional (EFA2) and all of the 32 items in the six-dimensional solution (EFA6) are within-item multidimensional. The Q-matrices constructed from the exploratory analyses are presented in Table 6.1.

Lastly, a three-dimensional Q-matrix representing test content and cognitive psychological theory is also employed in this study. This Q-matrix was constructed as part of an earlier research study (Gushta, Yumoto, & Williams, 2009) by assigning items to appropriate levels of the revised *Bloom's Taxonomy for Educational Objectives* (Anderson & Krathwohl, 2001; Bloom, 1956) which describe the cognitive processes necessary to successfully answer test items according to the Cognitive Process Dimension, independent of specific subject-area requirements. While there are six categories in the Cognitive Process Dimension, only 3 were represented in this assessment: *Remembering* (Factor 1; 3 items), which is the most basic cognitive process indicating that test items require only retrieval of stored information; *Understanding* (Factor 2; 14 items), a more complex process requiring summarizing and comparing; and *Application* (Factor 3; 15 items), for items requiring the use of procedures to solve familiar and novel tasks. Unlike the Q-matrices resulting from EFA solutions, the cognitive complexity Q-matrix (COG) follows simple-structure and the items are all between-item multidimensional. This Q-matrix is also presented in Table 6.1.

Table 6.1

Q-matrices Resulting from 2- and 6-Dimensional Exploratory Factor Analysis and Cognitive Complexity

Item	EFA2		EFA6						COG		
	1*	2	1	2	3	4	5	6	1	2	3
1	0	1	1	0	0	1	0	0	0	0	1
2	1	1	0	1	0	0	0	1	0	1	0
3	1	1	0	0	1	0	0	1	1	0	0
4	0	1	1	0	0	0	0	1	0	0	1
5	1	0	0	0	0	1	0	1	0	0	1
6	1	1	0	1	0	1	0	0	0	1	0
7	1	1	0	1	1	0	0	0	0	1	0
8	1	0	0	1	0	0	0	1	0	1	0
9	1	1	1	1	0	0	0	0	0	1	0
10	1	1	0	1	1	0	0	0	1	0	0
11	1	1	0	1	0	0	0	1	0	0	1
12	1	0	0	1	1	0	0	0	0	1	0
13	1	1	1	1	0	0	0	0	0	1	0
14	1	1	0	1	1	0	0	0	0	0	1
15	1	0	0	1	0	1	0	0	0	0	1
16	1	1	0	0	1	0	0	1	0	1	0
17	1	1	1	0	0	1	0	0	1	0	0
18	1	1	0	1	0	0	1	0	0	1	0
19	1	1	1	1	0	0	0	0	0	0	1
20	1	0	0	0	1	1	0	0	0	1	0
21	1	0	0	0	0	0	1	1	0	0	1
22	1	1	0	1	0	1	0	0	0	0	1
23	1	1	1	1	0	0	0	0	0	0	1
24	1	0	0	0	1	0	1	0	0	0	1
25	1	0	0	1	0	0	0	1	0	1	0
26	1	0	0	0	0	1	1	0	0	0	1
27	1	1	0	0	0	1	1	0	0	1	0
28	1	1	0	0	1	1	0	0	0	0	1
29	1	1	0	1	0	1	0	0	0	0	1
30	1	1	0	0	1	1	0	0	0	0	1
31	1	1	1	0	0	0	0	1	0	1	0
32	1	1	0	1	0	0	1	0	0	1	0

Note: Shaded entries indicate misfit according to the joint criteria; strikethrough indicates *Q*-matrix revision.

* Numbers denote latent factors.

6.3.Results

6.3.1 Original Models

Two-parameter normal-ogive (2-PNO) multidimensional item response theory (MIRT) models were fit for each of the EFA2, EFA6, and COG Q-matrices using Mplus version 6.11 (Muthén & Muthén, 1998-2010) and the specifications detailed in Chapter 3. MDIFF values in the simulation study portion of this dissertation were specified as low, moderate, and high difficulty and represent increasing discrepancy from the mean of the latent factor scores. As such, negative MDIFF values were not included but are hypothesized to affect fit indices as would positive MDIFF values of similar magnitude. Therefore, absolute MDIFF estimates and the minimum, mean, median, and maximum of such resulting from the real data analysis will be compared to the MDIFF values specified in the simulation design conditions to determine the corresponding item type.

Item parameter estimates for EFA2 are presented in Table 6.2 with minimum, mean, median, and maximum absolute MDIFF = [0.024, 1.006, 0.668, 8.918], respectively; estimated MDISC values range [0.124, 1.552] with a mean of 0.603; inter-factor correlation is estimated as $\hat{\rho}_{1,2} = -0.447$, and latent factor scores, or student ability, is distributed $\bar{\theta}' = [0.006, 0.038]$ with $\hat{\sigma}_{\theta} = [0.886, 0.725]$. These characteristics suggest that EFA2 approximates the moderate inter-factor correlation and moderate-discrimination / high-difficulty (i.e., large discrepancy between MDIFF and mean latent factor scores) conditions employed in the previous simulation study.

Table 6.2

Item Statistics Estimated for the 2-Dimensional Exploratory Factor Analysis Model

Item	MDIFF	MDISC	S- χ^2	MI		Wald	
				1	2	1	2
1	-8.918	0.149	30.605	4.361†			-1.824*†
2	-0.786	0.672	14.377			10.950	-0.863*†
3	-0.379	0.521	25.439			7.481	-3.071*†
4	-0.913	1.396	45.948*†	4.356†			-13.859*
5	-0.516	0.840	17.075		0.236	20.360	
6	0.529	0.268	18.671			4.347*	-0.940*†
7	-2.455	0.462	32.443			6.460	-2.029*†
8	-0.024	0.124	26.002		1.722	2.811*†	
9	-1.358	0.479	13.387			6.756	-2.734*†
10	-1.096	0.564	30.457			8.074	-2.743*†
11	-0.380	0.809	23.422			12.802	-2.295*†
12	0.326	0.467	21.542		0.766	10.538*	
13	-0.760	0.993	36.626*†			13.125	-3.658*†
14	-0.621	0.846	18.229			13.199	-1.115*†
15	-1.111	0.582	13.909		0.279	13.518*	
16	-0.901	0.460	40.203*†			6.900	-2.488*†
17	-2.032	0.486	32.176			4.826*	-4.155*
18	-1.853	0.321	19.577			4.914*	-1.413*†
19	-0.456	1.552	51.566*†			2.185*†	-8.964
20	1.453	0.701	32.349†		0.004	12.052*	
21	-0.087	0.706	25.340		0.030	16.694	
22	-0.765	0.559	30.872			8.735	-1.317*†
23	-0.714	0.402	14.542			6.125	-2.013*†
24	0.380	0.585	37.522*†		1.639	13.213*	
25	0.260	0.456	25.230		0.305	10.280*	
26	0.266	0.862	44.436*†		0.029	18.821	
27	-0.201	0.543	23.729			8.465	-2.240*†
28	1.057	0.542	14.683			8.231	-1.534*†
29	-0.195	0.653	24.413			9.954	-2.366*†
30	0.414	0.338	20.857			5.417*	-1.228*†
31	-0.731	0.505	23.979			4.471*	-5.506
32	-0.242	0.450	30.557			7.478	-0.946*†

*Note: Shaded entries indicate misfit according to the joint criteria.*** Misfit according to empirical cut point.**† Misfit according to theoretical cut point.*

Table 6.3 presents the item parameter estimates for EFA6 with minimum, mean, median, maximum absolute MDIFF values of [0.032,0.825,0.589,4.082]; MDISC values ranging [0.095,3.251] with a mean of 0.725; inter-factor correlations are estimated as:

$$\hat{\rho} = \begin{bmatrix} 1 & & & & & \\ 0.543 & 1 & & & & \\ 0.370 & 0.310 & 1 & & & \\ 0.482 & 0.433 & 0.832 & 1 & & \\ 0.435 & 0.821 & 0.520 & 0.607 & 1 & \\ 0.782 & 0.432 & 0.719 & 0.536 & 0.608 & 1 \end{bmatrix};$$

and latent factor scores are distributed $\bar{\theta}' = [-0.061, -0.006, 0.006, -0.002, 0.012, -0.006]$ and $\hat{\sigma}_{\theta} = [0.771, 0.850, 0.654, 0.771, 0.792, 0.783]$. Parameter estimates for the EFA6 model suggest that it approximates the moderate inter-factor correlation and moderate-discrimination / high-difficulty condition specified in the simulation study.

Table 6.3

Item Statistics Estimated for the 6-Dimensional Exploratory Factor Analysis Model

Item	MDIFF	MDISC	S- χ^2	MI						Wald					
				1	2	3	4	5	6	1	2	3	4	5	6
1	-4.082	0.333	45.128*†		0.006	2.897		5.203*†	0.003	3.202*†			-1.425*†		
2	-0.962	0.555	17.671	0.927		0.372	0.967	0.934			3.047*†				2.876*†
3	-0.335	0.649	27.494	3.142	0.726		0.257	0.064				2.320*†			3.661*
4	-0.770	2.343	22.650		0.002	2.798	1.111	1.296		6.909					-0.607*†
5	-0.604	0.742	29.852†	6.485*†	0.754	0.427		0.826					5.465		3.329*†
6	0.358	0.399	57.584*†	0.189		0.001		0.862	0.023		2.722*†		-0.569*†		
7	-2.146	0.534	35.246*†	1.037			2.838	1.042	0.538		5.679	0.317†*			
8	-0.032	0.095	41.875*†	2.086		0.016	0.572	0.097			0.764*†				0.447*†
9	-1.258	0.522	17.994			0.014	0.034	0.771	0.345	1.288*†	7.060				
10	-1.083	0.583	31.640†	0.400			0.029	0.808	0.275		4.668*	2.559*†			
11	-0.439	0.722	24.054	0.397		0.006	0.285	0.004			3.899*				3.382*†
12	0.376	0.406	28.580†	0.360			0.027	0.130	0.927		4.600*	1.608*†			
13	-0.680	1.175	34.731*†			0.253	0.895	0.720	0.175	1.005*†	13.247				
14	-0.668	0.800	18.591	0.335			1.083	0.237	0.489		6.631	2.355*†			
15	-1.336	0.496	15.502	0.127		0.118		0.248	0.052		1.372*†		2.451*†		
16	-0.491	0.947	43.118*†	1.715	0.771		1.106	0.057				-1.242*†			5.050*
17	-1.852	0.55	34.096*†		0.940	0.000		0.709	1.819	3.675*†			5.038*		
18	-2.052	0.291	22.526	0.655		0.237	0.003		0.783		1.840*†			0.921*†	
19	-0.528	1.154	27.618†			0.398		0.246	1.709	6.409	2.892*†				
20	1.500	0.721	26.863†	0.592	0.054				0.379			4.080*	3.003*†		
21	-0.107	0.575	32.303†	0.505	1.199	2.142	0.065							4.645*	3.191*†
22	-0.573	0.766	56.644*†	0.767		0.243		0.376	1.626		4.858*		-0.720*†		
23	-0.650	0.445	22.087			0.252	1.824	2.732	0.232	0.852*†	6.602				
24	0.422	0.539	38.982*†	2.207	0.669		0.042		1.221			2.681*†		4.558*	
25	0.333	0.358	25.456	1.288		0.143	0.077	0.144			1.884*†				2.325*†
26	0.117	3.251	37.679*†	0.035	0.003	3.050			0.130				-1.112*†	2.425*†	
27	-0.213	0.541	36.953*†	3.769*	0.056	0.002							1.986*†	2.592*†	
28	0.954	0.630	16.552	1.205	0.064			0.471	2.993			2.105*†	5.902		
29	-0.220	0.599	22.847	1.478		2.233		0.798	2.151		2.875*†		2.111*†		
30	0.280	0.537	17.389	0.335	0.684			0.203	0.832			4.903*	0.632*†		
31	-0.695	0.547	26.066		0.444	0.427	2.879	1.185		3.651*†					4.960*
32	-0.271	0.406	29.925†	0.001		1.713	0.003		2.091		1.133*†			2.505*†	

Note: Shaded entries indicate misfit according to the joint criteria.

* Misfit according to empirical cut point.

† Misfit according to theoretical cut point.

Parameter estimates for the final model, COG, are presented in Table 6.4. Absolute values of the MDIFF minimum, mean, median, and maximum are [0.025, 1.179, 0.612, 15.988], respectively; the MDISC values range 0.082 to 1.148 with a mean of 0.594; inter-factor correlations are estimated as $\hat{\boldsymbol{\rho}} = \begin{bmatrix} 1 & & \\ 0.918 & 1 & \\ 0.940 & 0.981 & 1 \end{bmatrix}$; and latent factor scores are distributed $\bar{\boldsymbol{\theta}}' = [0.003, 0.003, 0.004]$ and $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\theta}} = [0.862, 0.899, 0.903]$. Given these estimates, the COG model approximates the highly-correlated, moderate-discrimination / high-difficulty condition from the simulation design.

Table 6.4

Item Statistics Estimated for the Cognitive Complexity Model

Item	MDIFF	MDISC	S- χ^2	MI1	MI2	MI3	Wald 1	Wald 2	Wald 3
1	-15.988	0.082	33.488†	4.814†	2.404				1.340*†
2	-0.758	0.694	19.310	0.053		0.024		16.663	
3	-0.309	0.653	75.387*†		0.446	0.601	12.992*		
4	-1.32	0.679	20.011	0.144	0.516				16.593
5	-0.535	0.790	17.693	0.012	0.282				19.806
6	0.477	0.297	19.609	0.049		0.039		6.844*	
7	-2.130	0.534	32.856†	3.385		6.054*†		11.073*	
8	-0.025	0.119	35.350†	1.306		1.497		2.748*	
9	-1.162	0.561	13.211	2.262		0.069		13.184*	
10	-0.903	0.703	87.309*†		1.941	3.861	14.088*		
11	-0.347	0.886	24.913	0.026	0.305				22.400
12	0.334	0.453	22.019	0.203		0.113		10.454*	
13	-0.666	1.148	56.987†	0.895		0.276		27.220	
14	-0.600	0.867	18.354	0.020	0.782				21.728
15	-1.148	0.558	16.650	3.600	0.134				13.403*
16	-0.777	0.534	38.607†	0.113		0.560		12.800*	
17	-1.667	0.594	64.357*†		3.936†	5.947*†	11.180*		
18	-1.630	0.365	18.206	0.582		0.433		8.454*	
19	-0.517	0.961	22.197	2.234	2.100				24.110
20	1.495	0.672	34.368†	1.346		2.894		11.842*	
21	-0.090	0.672	26.402	0.937	0.805				16.514
22	-0.711	0.601	32.389†	0.067					14.446*
23	-0.624	0.460	18.699	0.230	0.714				10.943*

Item	MDIFF	MDISC	S- χ^2	MI1	MI2	MI3	Wald 1	Wald 2	Wald 3
24	0.394	0.558	40.234†	2.158	4.051†				12.991*
25	0.267	0.442	25.785	0.041		0.018		10.184*	
26	0.276	0.813	43.706†	1.425	1.948				18.509
27	-0.177	0.615	25.229	1.492		2.952		14.455*	
28	0.967	0.593	21.334	0.475	0.032				12.913*
29	-0.175	0.729	24.254	1.360	0.011				17.555
30	0.375	0.372	29.002	2.309	0.117				8.567*
31	-0.653	0.553	21.738	1.624		1.522		13.098*	
32	-0.227	0.479	29.504	1.889		0.186		11.396*	

Note: Shaded entries indicate misfit according to the joint criteria.

** Misfit according to empirical cut point.*

† Misfit according to theoretical cut point.

The design-appropriate cut points for each model- and item-fit index are selected as the empirical cut points calculated from those simulated true model conditions that closely approximate the characteristics of the EFA2, EFA6, and COG models presented in Table 6.5. While cut points for six-dimensional models cannot be directly obtained from the results of the simulation study, none of the model-fit indices demonstrated sensitivity to number of dimensions, similar to the findings of Jackson (2007); therefore, the cut points for the three-dimensional model were employed with EFA6.

Table 6.5
Design Appropriate Cut Points for the Grade 6 Mathematics Achievement Real-Data Analysis

	Fit Statistic	EFA2		EFA6		COG
		B	W	B	W	B
Model	χ^2/df		1.074		1.073	1.073 (1.070)
	RMSEA		0.009		0.009	0.009 (0.008)
	GDDM		0.004		0.004	0.004 (0.004)
Item	S- χ^2	37.519	33.740	35.844	32.718	40.250 (41.466)
	MI	9.061		3.472	3.300	5.086 (6.325)
	Wald	15.284	5.626	12.883	5.438	13.498 (14.887)

Model-fit for the EFA2 model is estimated as $\chi^2/\text{df} = 1.205$, RMSEA = 0.014, and GDDM = 0.012 which suggests model misfit for all three indices according to the empirical cut points but does not suggest misfit according to the theoretical cut points ($\chi^2/\text{df} = 2.0$; RMSEA = 0.05). Model-fit values and cut points differ slightly for the EFA6 model: $\chi^2/\text{df} = 1.099$, RMSEA = 0.010, and GDDM = 0.006 suggesting misfit for all three indices but again does not suggest model misfit under the theoretical cut points. The COG model demonstrates the worst fit overall as $\chi^2/\text{df} = 1.490$, RMSEA = 0.022, and GDDM = 0.007 which also suggests misfit according to all three model-fit indices but, as with the previous models, does not demonstrate misfit under the theoretical cut points.

Noting that the model-fit indices generally reject these three models as misspecified, lessons learned from the simulation study presented in this dissertation can be applied to the examination of specific item-fit results for the purpose of model revision and Q-matrix amendment as follows.

The simulation study in this dissertation suggested that, under conditions similar to those of the real data analysis, the $S-\chi^2$ has poor-to-moderate power overall to predict misspecified items when the model has been identified as misspecified, the Modification Indices have moderate-to-strong power for the EFA2 model and poor-to-moderate power for EFA6 and COG, and Wald Tests have poor power for the EFA2 and COG models and poor-to-moderate power for the EFA6 model. Additionally, of the three item-fit indices only the Modification Index demonstrated sensitivity to number of dimensions and, therefore, requires special consideration in application to EFA6. To account for this sensitivity, Modification Index cut points for six dimensions were extrapolated based on the ratio of the values observed for the two- and three-dimensional models. These values were calculated separately for simple- and complex-structure models and presented in Table 6.5 along with all other model- and item-fit cut points.

Further, previous research has suggested that model revision according to Modification Indices when misspecification is severe resulted in poor recovery of the true population model (Hutchinson, 1998) while the Wald Test performed well in identifying misspecified parameters when guided by theoretical justification (Chou & Bentler, 2002). The Modification Indices and Wald Test statistics indicate direct or implied change in overall model fit should a particular parameter be freed or fixed; thus, whenever these statistics indicated multiple parameter revisions, only the most strongly indicated revision

was considered. For example, given multiple significant MI values, the largest MI value will be selected for revision; with multiple Wald Test statistics indicating misfit, the value closest to zero is selected for use in model revision. Joint criteria for identifying misfit using the three types of item-fit indices are, therefore, defined as requiring a significant $S-\chi^2$ and a significant MI or Wald Test value – the $S-\chi^2$ results providing a conservative limitation to the number of statistically-determined model revisions. The following revisions of the three Q-matrices are suggested according to the joint criteria.

Evaluating item-fit for the EFA2 model according to the empirical cut points leads $S-\chi^2$ to reject six of 32 items, MI1 and MI2 to reject none of the items, Wald Test 1 to reject 12 of 30 items since only items loading on factor 1 are eligible for this statistic, and Wald Test 2 to reject 21 out of 23 items. When the item-fit results are considered jointly, the combination of $S-\chi^2$ and either the Modification Index or Wald Test indicates that the Q-matrix entries for Item 13 should be re-specified as $Q_{13,(1,2)} = [1, \mathbf{0}]$, Item 16 re-specified as $Q_{16,(1,2)} = [1, \mathbf{0}]$, and Item 19 as $Q_{19,(1,2)} = [\mathbf{0}, 1]$, where the bolded Q-matrix elements indicate deletion based on the joint information provided by the $S-\chi^2$ and Wald Test fit values (see also Table 6.1). Although Items 4 and 24 are indicated as misspecified by the joint criteria, these items are not re-specified since the Wald Test results suggest deleting the only Q-matrix entry for those items. Were the theoretical cut points employed, five items would be indicated as misfitting overall with the Modification Indices over-identifying misfit and the Wald Test statistics under-identifying misfit.

When the EFA6 model is estimated, 11 of the 32 items are identified as misfitting according to the empirical $S-\chi^2$ cut points; two items are identified as misfitting by MI1, one item was identified as misfitting according to MI5, and none were identified as

misfitting by MI2, MI3, MI4, and MI6. The Wald Tests indicated that six to 13 of the items were misfitting. Combining this evidence according to the joint criteria described earlier, we can conclude that 11 items demonstrate misfit; the suggested revisions are presented in Table 6.1. Had the theoretical cut points been used, 18 items would have been identified as misspecified by the joint criteria.

Lastly, the COG model demonstrated the worst overall model-fit but the best overall item-fit. Five misfitting items were identified as misfitting by the $S-\chi^2$ index, two items were identified as misfitting by the MI3 index, three items were identified as misfitting by the Wald Test 1, 12 items were identified as misfitting by the Wald Test 2, and seven items were identified as misfitting by the Wald Test 3. The result is that only 3 misfitting items are identified according to the joint criteria. According to the theoretical cut points, four items would be indicated as misfitting.

As shown in Table 6.1, the misfitting items for the COG model are suggested to be re-specified as $Q_{3,(1,2,3)} = [0, 0, 0]$, $Q_{10,(1,2,3)} = [0, 0, 0]$, $Q_{17,(1,2,3)} = [0, 0, 1]$. These results suggest that all Q-matrix entries associated with the first factor, *Remembering*, be deleted. Taking this into consideration the COG model is re-specified as a two-dimensional model; Item 3 and Item 10 are subsequently associated with latent factor 3, *Application*, based on the largest MI value. Interpreting this revision with respect to Bloom's Taxonomy, these items are suggested to require higher-order cognitive operations than originally presumed; items categorized as *Remembering* which were not part of topics delivered directly via instruction would result in higher cognitive demands than originally anticipated. The suggested re-specification for Item 17 can be interpreted with respect to levels of Bloom's Taxonomy as suggesting that the item requires the

cognitive operations associated with *Application* instead of *Understanding*, again suggesting higher-order cognitive processing. Examination of the test content could show these to be a reasonable re-specifications of the Q-matrix.

6.3.2 Revised Models

A variety of suggestions were made for the revision of the EFA2, EFA6, and COG Q-matrices in the previous section. Since the COG model was reduced from a three-dimensional to two-dimensional model additional, appropriate, cut points are provided in parentheses in Table 6.5; otherwise, the same empirical cut points are applied to the model- and item-fit estimates resulting from estimation of models according to the revised Q-matrices. These revised Q-matrices were constructed and the models re-estimated. The resulting model-fit estimates are presented in Table 6.6 against those resulting from the original Q-matrices, revealing a complicated picture. While all models continue to demonstrate misfit, fit of EFA2 worsens according to the χ^2/df and RMSEA but improves according to the GDDM; fit of EFA6 worsens according to the χ^2/df and RMSEA but remains the same according to the GDDM; and fit of the COG model improves according to the χ^2/df while staying the same for the RMSEA and GDDM.

Table 6.6
Model-Fit Estimates for the Original and Revised Models

Model	Statistic	Original	Revised
EFA2	χ^2/df	1.205	1.229
	RMSEA	0.014	0.015
	GDDM	0.012	0.010
EFA6	χ^2/df	1.099	1.110
	RMSEA	0.010	0.011
	GDDM	0.006	0.006
COG	χ^2/df	1.490	1.486
	RMSEA	0.022	0.022
	GDDM	0.007	0.007

Table 6.7 presents the item-fit results for the revised EFA2 model, for which the Q-matrix entries for items 13, 16, and 19 were modified as shown in Table 6.1. As a result of these revisions, the S- χ^2 now identifies five items as misfitting (six were identified in the original model), no items are identified as misfitting according to the Modification Indices (similar to the original model), 12 items are identified as misfitting by Wald Test 1 and 19 by Wald Test 2 (previously 12 and 21). The joint criteria indicate that the Q-matrix entries for two items should be additionally revised; indicated as misfitting under the original model, the Wald Test results suggest that Q-matrix entries associating items 16 and 24 with latent factor 1 be deleted. However, these items would then be unassociated with any latent factor, therefore, these revisions are not advised.

Table 6.7
Item-Fit Values for the Revised EFA2 Model

Item	S- χ^2	MI		Wald	
		1	2	1	2
1	30.768	4.377			-1.695*
2	13.728			9.864	0.020*
3	25.103			6.499	-2.573*
4	60.341*				-20.479
5	18.983		2.533	20.175	
6	18.283			3.455*	-1.124*
7	32.130			6.163	-1.041*
8	26.447		2.176	2.788*	
9	13.179			6.302	-1.653*
10	30.750			6.945	-2.433*
11	25.031			11.912	-1.022*
12	23.083		0.078	10.504*	
13	32.104		7.518	27.780	
14	19.770			11.772	-0.252*
15	14.723		0.085	13.455*	
16	44.114*		2.572	12.859*	
17	31.993			4.490*	-3.200*
18	19.281			4.117*	-1.291*
19	129.910*	0.498			-24.699
20	35.444		0.204	12.008*	
21	26.861		0.523	16.622	
22	29.621			7.858	-0.752*
23	14.909			5.512	-1.466*
24	38.862*		3.396	13.134*	
25	26.801		0.102	10.276*	
26	48.559*		0.177	18.652	
27	24.270			7.458	-1.793*
28	15.323			7.198	-1.147*
29	25.264			8.996	-1.696*
30	21.244			4.570*	-1.119*
31	26.540			3.973*	-4.700*
32	32.040			6.744	-0.347*

The item-fit results for the revised EFA6 model, presented in Table 6.8, present a picture as complicated as the original model. The $S-\chi^2$ statistic indicates that 14 items are misspecified, the Modification Indices suggest a total of 16 revisions, and the Wald Tests suggest 47 revisions; as compared to 11, 3, and 55 revision suggestions under the original model. Further, the joint criteria suggest that 11 items are candidates for revision – the same number and a certain degree of overlap with the item-fit results evidenced under the original model (8 items). The number of Q-matrix revisions suggested by these results is greater than can be reasonably described within the scope of this dissertation; therefore specific recommendations are not presented. These results do indicate that as Q-matrix entries are deleted via the Wald Test results, Modification Indices suggest alternate associations between items and latent factors. Additional iterations of Q-matrix and item-fit evaluation and revision appear to be necessary to achieve a point of stability in which revisions are no longer necessary or possible.

Table 6.8
Item-Fit Values for the Revised EFA6 Model

Item	$S-\chi^2$	MI						Wald					
		1	2	3	4	5	6	1	2	3	4	5	6
1	82.828*		1.365	6.326*	3.944*		0.016	4.175*				-3.081*	
2	18.601	1.080		0.114	3.607*	0.653			3.461*				1.835*
3	30.315	1.476	0.117		0.450	2.230				3.661*			1.756*
4	22.909		12.274*	6.826*	6.208*	6.080*		7.984					-1.943*
5	39.167*	3.299	2.576	0.232		0.674					4.827*		4.147*
6	35.752*	0.635		0.215	0.130	0.021	0.025		6.871*				
7	37.464*	0.138			2.514	0.154	0.340		11.162*				
8	51.117*	2.456		0.026	1.437	0.769			2.775*				
9	19.855			0.004	0.192	1.131	0.029	0.608*	4.556*				
10	29.025	0.182			0.000	0.263	0.731		2.608*	2.945*			
11	23.652	0.393		0.184	0.432	0.065			3.275*				2.888*
12	28.920	2.478			2.040	0.795	1.867		3.537*	1.246*			
13	35.523*	0.494		0.008	0.003		0.009		27.192				
14	17.428	1.661			0.003	0.398	1.055		5.550	1.948*			
15	16.887	0.009		0.737		0.691	0.347		2.956*		2.121*		
16	45.574*	3.023	0.138	0.387	0.922	0.354							11.266*
17	34.899*		14.685*	3.496*	6.709*	6.857*	9.542*	12.517*					
18	22.744	1.643		0.060	0.088		1.485		1.271*			1.453*	
19	32.851*				0.484	1.005	6.144*	5.808	-0.755*				
20	29.325	1.242	0.509			0.514	0.722			3.578*	2.939*		
21	24.964	0.133	0.115	0.533	1.611							4.595*	2.715*
22	34.148*	0.354		0.373	0.026	0.401	1.034		14.432				
23	22.650			0.025	0.543	0.991	0.003	0.884*	3.871*				
24	37.479*	0.108	1.245	3.941*	0.174		4.450*					13.446*	
25	25.969	1.051		0.001	0.086	0.002			1.391*				2.215*
26	53.005*	0.005	0.011	0.356	0.709		0.019					18.288	
27	35.792*		0.396	0.553	2.107		2.254	2.301*				6.262	
28	20.271	2.960	0.144			1.634	2.199			1.690*	4.604*		
29	24.633	2.391		3.190		0.369	1.488		5.658		1.651*		
30	18.034	0.066	0.278			1.499	0.290			5.068*	-0.476*		
31	24.863		2.203	0.003	3.848*	1.620		4.268*					2.639*
32	35.236*	0.103		1.587	0.006		1.587		1.034*			2.606*	

Fewer revisions to the COG model were suggested according to the joint criteria than for either the EFA2 or EFA6 model. After making the three suggested revisions, collapsing the model to two latent factors, and estimating the model according to the revised Q-matrix, the $S-\chi^2$ identifies two items as misfitting, the Modification Indices do not identify any items as misfitting, and the Wald Tests identify 21 items. The original model had identified 5 items, 2 items, and 22 items as misfitting. No items are identified as misfitting according to the joint criteria (Table 6.9) when the revised COG model is estimated.

Table 6.9
Item-Fit Values for the Revised COG Model

Item	S- χ^2	MI		Wald	
		1	2	1	2
1	32.406	2.008			1.342*
2	29.355		0.014	16.662	
3	30.762	0.168			14.846*
4	19.358	0.511			16.603
5	18.069	0.342			19.876
6	16.937		0.037	6.844*	
7	37.829		5.798	11.070*	
8	29.555		1.414	2.747*	
9	18.400		0.016	13.188*	
10	31.481	0.072			15.883
11	25.693	0.428			22.394
12	20.632		0.092	10.452*	
13	81.323*		0.174	27.217	
14	18.167	0.730			21.771
15	15.929	0.225			13.405*
16	41.456		0.660	12.802*	
17	33.557				12.651*
18	18.250		0.388	8.452*	
19	22.339	1.619			24.185
20	32.486		2.960	11.840*	
21	26.527	0.912			16.535
22	31.132				14.442*
23	17.173	0.886			10.946*
24	39.015	3.935			12.980*
25	24.081		0.013	10.183*	
26	44.602*	1.958			18.487
27	29.801		2.916	14.455*	
28	19.625				12.901*
29	24.115				17.553
30	26.128	0.041			8.571*
31	26.891		1.409	13.099*	
32	30.381		0.094	11.393*	

6.4. Summary

This real data analysis demonstrates the usefulness in considering the psychometric properties of items and models as well as sample characteristics of the assessment data when examining model- and item-fit for the purpose of evaluating the Q-matrix. The use of fit index cut points appropriate for the number of latent factors, sample size, test length, strength of inter-factor correlations, item multidimensionality, and the broad classifications of item discrimination and difficulty to jointly consider model- and item-fit information identified a manageable number of model revisions. Use of the suggested or theoretical cut points, however, leads to dissonant results as the model-fit statistics would suggest that all three models fit the data while the item-fit statistics would generally over-identify item misfit.

Applying the empirical, design-appropriate, cut points in a single iteration of model criticism and evaluation, the three Q-matrices were re-specified according to the joint information provided by the item-fit indices. Model-fit information resulting from these Q-matrix re-specifications does not clearly indicate overall improvement or worsening of model-fit. Item-fit information, however, does suggest that correctly specified Q-matrices can be obtained through such an iterative re-specification process. Information provided by the initial model estimation and first iteration of revisions indicate consistent and reasonable results.

Estimation of the EFA2 model, which represented a general under-factoring of the model vis-à-vis the best-fitting dimensionality structure suggested by exploratory analysis, resulted in a degree of model- and item-misfit which could be improved by Q-matrix edits for a modest number of misfitting items. Revision of the Q-matrix and re-

estimation of the model resulted in fewer items flagged as misspecified and fewer items identified as candidates for revision, though it should be noted that some of the final suggested revisions are not feasible or require item deletion.

The EFA6 model, which estimates the number of dimensions suggested by exploratory analysis, demonstrates some degree of model misfit as well as the greatest degree of item misfit and largest number of suggested Q-matrix revisions. The item-fit results for the original Q-matrix suggested numerous deletions of Q-matrix entries as a result of the Wald Test statistic values; item-fit results subsequent to these edits, however, indicate an increased number of additions to the Q-matrix as suggested by the Modification Indices. While it was not in the scope of this study to iterate the Q-matrix revision to a point of stability with regards to the item-fit results it is apparent that the fit statistics are suggesting modest and reasonable restructuring of the Q-matrix and not simply attempting to build a saturated model.

Finally, the COG model, specified according to theory, yielded the worst overall model fit but the fewest overall revisions of the Q-matrix according to the joint information provided by the item-fit results. Upon making these edits, overall item misfit was greatly reduced; the joint criteria used to identify candidate items for further revision failed to identify further misfitting items. In a single revision, the Q-matrix achieved stability with respect to item-fit information. While these results may suggest that the Q-matrix resulting from the COG model is a candidate for the correct, or true, Q-matrix, there are two additional considerations that must be noted. First, the final inter-factor correlation for the revised COG model is $r = 0.98$, suggesting that the model is actually unidimensional, though this concern can be disputed on the grounds of evidence

presented in previous research (Adams & Wu, 2002; Wu & Adams, 2006). Second, the Wald Test statistics continue to suggest that a number of Q-matrix entries be deleted, which can be understood by considering the MDISC values. Shown to be sensitive to MDISC in Chapter 4, the Wald Test values are nearly perfectly positively correlated with the MDISC estimates, suggesting that the weak-to-moderate discrimination of items estimated by this model is directly contributing to the identification of item misfit.

At the conclusion of this first iteration of model revision, it must be noted that all of the models continue to demonstrate misfit even though the Q-matrix has been revised. While some of the model-fit indices show improvement, others do not and this is especially true for the COG model. The fact that model improvement is suggested by the item-fit indices but fails to materialize when models are revised and re-estimated suggests that the Q-matrices may represent random patterns of associations which would not be expected to appropriately capture variability in the model. Final acceptance of any of the Q-matrices presented in the real data analysis portion of this dissertation would require further analysis and substantive consideration, beyond the scope of the current study.

The task of evaluating and revising Q-matrices when applied to real data is further complicated by the fact that the true or correct Q-matrix is unknown and, therefore, its recovery cannot be directly evaluated. It is in fact possible that any number of equivalent models could result in the estimation of vastly different parameters but the same sets of statistical fit indices (Raykov & Marcoulides, 2001); alternately, it is possible that mathematically equivalent Q-matrices exist for an accepted Q-matrix (Bechger, Verstralen, & Verhelst, 2002). Raykov and Marcoulides (2001) state that, since statistical

indices do not exist which can distinguish equivalent models, model selection under such conditions must be managed by substantive consideration.

The use of empirically-derived model- and item-fit cut points yields results demanding thoughtful and careful consideration of the elements of these three different Q-matrices, for which the first round of model criticism and revision has been presented. Had the theoretical cut points been employed to evaluate model- and item-fit they would have first suggested that the overall model fit the data well, likely deterring further model criticism which would then have been complicated by inflated counts of misfitting items, as suggested by the inflated Type-I error rates presented in Chapter 4. Rich, appropriate, statistical information as provided by the multiple fit indices employed in this study serves to facilitate the decisions required of practitioners and researchers during the process of model evaluation.

Chapter 7

Discussion

The current study extends research on model- and item-fit sensitivity to consider the influence of item type, defined jointly according to item discrimination and item difficulty, and model misspecification via Q-matrix elements. Specifically the performance of three model-fit indices (χ^2/df , RMSEA, and GDDM) and three item-fit indices (S- χ^2 , Modification Index, and Wald Test) was investigated in a simulation study manipulating item type and degree of model misspecification as well as sample size, number of observed variables (test length), item multidimensionality (simple or complex factor structure), the number of latent factors, and the strength of the correlation between latent factors. These fit indices are typically available within either a confirmatory factor analysis (CFA) framework or multidimensional item response theory (MIRT) framework. Equivalence between models estimated within these two frameworks, however, is achieved by satisfying specific assumptions and parameter constraints, detailed in previous research (Kamata & Bauer, 2008; Takane & de Leeuw, 1987), providing researchers and practitioners with additional information in the evaluation of model performance and validity.

This chapter begins with a summary of key findings from the study. The original research questions focused on the distributional forms of the fit indices under true model estimation conditions, the sensitivity of the fit indices under true and misspecified model estimation, and the influence of simulation conditions on power rates for each model- and item-fit index. The results are, therefore, summarized with these points in mind. A discussion of its limitations and suggestions for future research follows.

7.1. Summary of Key Findings

The first investigation in this dissertation is an examination of the distributional forms of the model- and item-fit indices. The distributional forms of five of the six fit indices included in this study have been described according to known distributions; no distributional form of the generalized dimensionality discrepancy measure (GDDM; Levy & Svetina, 2010) has been defined. The χ^2/df ratio and RMSEA fit indices are stated to follow rescaled χ^2 distributions (Browne and Cudeck, 1993; Steiger, 2000; Steiger and Lind, 1980) with degrees of freedom defined as the model degrees of freedom; the S- χ^2 (Orlando & Thissen, 2000, 2003; Zhang & Stone, 2008) is χ^2 -distributed with degrees of freedom equal to the number of valid total score categories adjusted for the number of item parameters; and values of the Modification Index (Sörbom, 1989) and Wald Test (Buse, 1982) are evaluated as being χ^2 -distributed with a single degree of freedom. Previous research has suggested cut points of $\chi^2/\text{df} = 2$ or 3 (Byrne, 1989; Carmines & McIver, 1981; Hu & Bentler, 1999; Marsh & Hocevar, 1985) and RMSEA = 0.05 or 0.06 (Hu & Bentler, 1999) while cut points for the item-fit indices have been defined according to the critical values corresponding to a nominal significance level of $\alpha = 0.05$. The empirical cumulative distribution functions and measures of sensitivity, η^2 , resulting from estimation of the true models, however, indicate that these indices do not strictly adhere to the proscribed distributions and vary according to many of the conditions manipulated in this study. Further, many of the suggested cut points were determined based on descriptive analysis of model fit (e.g., Hu & Bentler, 1999), not inferential methods. Therefore, the 95th percentiles calculated in the current study were employed as

cut points, allowing for explicit model- and item-fit testing in subsequent analysis of misspecified models and items.

Summarizing the behavior of the model- and item-fit indices according to the various simulation conditions provides an interesting and complex picture. Key findings for the model- and item-fit indices according to the simulation conditions manipulated in this study are reviewed below. Additionally, Table 7.1 provides a quick reference indicating the conditions for which estimated models demonstrated the best fit under true model estimation, worst fit under misspecified model estimation, and the highest power rates for each of the fit indices.

Table 7.1:

Summary of Model and Item Fit Statistic Behaviour by Model Characteristics and Test Design Specifications

Fit Statistic	Interpretation	Number of Dimensions	Test Length	Sample Size	Multi-Dimensionality	Inter-Factor Correlation	Item Type	Misspecification
χ^2/df	Best fit:		Longer tests				Interacts with sample size: for small samples, lower item discrimination; for large samples, higher item discrimination	
	Worst fit:			Larger sample sizes		Weaker correlations	Higher item discrimination; well-targeted item difficulty	
	Best detection:		Longer tests	Larger sample sizes		Weaker correlations		
RMSEA	Best fit:		Longer tests	Larger sample sizes				
	Worst fit:	Fewer dimensions				Weaker correlations	Higher item discrimination; well-targeted item difficulty	
	Best detection:		Longer tests	Larger sample sizes		Weaker correlations		
GDDM	Best fit:		Longer tests	Larger sample sizes			Higher item discrimination; well-targeted item difficulty	
	Worst fit:	Fewer dimensions				weak correlation	Moderate item discrimination; well-targeted item difficulty	
	Best detection:		Longer tests	Larger sample sizes		weak correlation		
$S-\chi^2$	Best fit:		Shorter tests	Smaller sample sizes		Weaker correlations		
	Worst fit:	Fewer dimensions				Stronger correlations		Alternate-factoring (moderate misspecification); under-factoring (severe misspecification)
	Best detection:	Fewer dimensions		Larger sample sizes			Higher item discrimination; well-target item difficulty	

Fit Statistic	Interpretation	Number of Dimensions	Test Length	Sample Size	Multi-Dimensionality	Inter-Factor Correlation	Item Type	Misspecification
Modification Index	Best fit:	More dimensions		Smaller sample sizes		Stronger correlations		
	Worst fit:	Fewer dimensions		Larger sample sizes		Weaker correlations		
	Best detection:	Fewer dimensions	Shorter tests	Larger sample sizes		Weaker correlations		Under-factoring (severe misspecification)
Wald Test	Best fit:		Longer tests		Between-item multi-dimensionality (simple structure)		Higher item discrimination; well-targeted item difficulty	
	Worst fit:			Smaller sample sizes			Moderate item discrimination; poorly-targeted item difficulty	Under-factoring (severe misspecification)
	Best detection:	More dimensions				Weaker correlations	Moderate item discrimination; well-targeted item difficulty	Alternate-factoring (moderate misspecification)

Very few effects due to *simulation condition 1: number of dimensions* were observed, none of them for model-fit indices. These results conform with that of previous research which showed that values of the χ^2/df and RMSEA are not sensitive to the number of latent factors in a misspecified model (Beauducel & Wittman, 2005; Jackson, 2007). Looking back to the formulas for the model-fit indices, it can be seen that the number of latent factors are not directly included, with the exception of the GDDM. The result is that any effect of the number of latent factors appears only indirectly through other parameters. Values of the Modification Indices under true model estimation are seen to increase with number of dimensions, suggesting better fit. The $S\text{-}\chi^2$, however, demonstrates decreased power to detect misfitting items as the number of factors increases.

Many fit indices demonstrated sensitivity to *simulation condition 2: test length* when true models were estimated. All of the model-fit indices as well as the Wald Test statistics demonstrated improved fit for true models as test length increased. Hu and Bentler (1999) and Jackson (2007) both reported power rates for the RMSEA that increased with test length. This effect could be anticipated as an increase in the number of observed variables corresponds to an increase in overall precision when the variables are of high discriminatory power. Values of the $S\text{-}\chi^2$, however, increase with test length, indicating worse fit, which also corresponds to the findings of Zhang and Stone (2008). One component of the $S\text{-}\chi^2$ is the joint likelihood of all possible response patterns which increases with every additional item. Further investigation is necessary is required to determine if this directly corresponds to power rates.

An effect of *simulation condition 3: sample size* is present across many of the fit indices. Under true model estimation, the values of the RMSEA and Wald Test statistics decrease with sample size. Previous research has suggested that increased sample size results in decreased sampling variability and, therefore, improved model fit (e.g., Beauducel & Wittman; Hu & Bentler, 1999; Jackson, 2007). The values of the GDDM, $S-\chi^2$, and Modification Indices, however, increase with sample size, denoting worsened fit. These fit statistics fail to explicitly incorporate sample size in their calculations and may benefit from sample size adjustment. When misspecified models are estimated, power increases with sample size for the χ^2/df , RMSEA, $S-\chi^2$, and Modification Indices. This is aligned with previous research that showed the χ^2/df (Marsh, Hau, & Wen, 2004) and RMSEA (Beauducel & Wittman, 2005; Curran et al., 2003; Fan & Sivo, 2005, 2007; Fan, Thompson, & Wang, 1999; Sivo, Fan, Witte, & Willse, 2006) to be modestly sensitive to sample size and the power rates of the $S-\chi^2$ and Modification Indices to increase with sample size (Hutchinson, 1998; Zhang & Stone, 2008).

None of the model-fit indices demonstrated sensitivity to *simulation condition 4: multidimensionality* under true model estimation. The Wald Test statistics, however, were shown to worsen for within-item multidimensionality. Since the Wald Test is calculated as the ratio of a factor loading to its standard error, these results suggest that within-item multidimensionality contributes to imprecision of parameter estimates. Under misspecified model estimation, power rates calculated according to the GDDM are seen to increase for complex-structure models, where items demonstrate within-item multidimensionality. Previous research by Fan and Sivo (2005, 2007) and Hu and Bentler (1998) found that model fit according to the χ^2/df and RMSEA fit indices worsened for

models estimated as under-factored. With fewer estimated parameters, the remaining parameters are increasingly subject to sampling variability and, therefore, likely to result in misfit.

Simulation condition 5: inter-factor correlation also demonstrated little effect on model-fit indices when true models were estimated. Stronger correlations, however, corresponded to worse fit for the $S\text{-}\chi^2$ and better fit for the Modification Indices. Power rates for the χ^2/df , RMSEA, Modification Indices, and Wald Test statistics were all highest when inter-factor correlation was weak. These results correspond to those found by Ximénez (2009) who reported that RMSEA values decreased for misspecified models when factors were moderately correlated versus uncorrelated.

Finally, *simulation condition 6: item type* showed very little effect on the majority of the fit indices when true models were estimated, though fit according to the Wald Test statistics improved with larger MDISC values and worsened as MDIFF became increasingly discrepant from the mean of the latent factor distribution. Under model misspecification, however, effects of item type on power rates appear. The GDDM correctly rejects misspecified models at higher rates when items are both highly discriminating and highly discrepant from the latent factor distribution, the $S\text{-}\chi^2$ demonstrates the highest power when items are also highly discriminating but well-targeted to the latent factor distribution, and the Wald Test statistics demonstrate power rates that are higher for moderately-discriminating, well-targeted items. There is some precedent for the effect of MDISC in the literature: Beauducel and Wittman (2005) and Jackson (2007) showed that the χ^2/df and RMSEA demonstrated sensitivity to indicator reliability, where higher reliabilities resulted in larger fit values which correctly

indicated misspecification. Indicator reliabilities, or factor loadings, and MDISC differ as a matter of a known transformation, therefore, these results are similar and applicable to the results shown in this dissertation.

Notably missing from the above descriptions of fit statistic performance is the effect of model misspecification. Degree of model misspecification, moderate or severe, was associated with very small percentages of variance in the model-fit statistics indicating little to no effect. These results differ from previous studies which showed the RMSEA and χ^2/df to indicate worse fit as degree of misspecification increased (Fan & Sivo, 2005, 2007; Jackson, 2007; Ximénez, 2009). Modification Indices and Wald Test statistics, however, were found to be sensitive to the specific types misfit introduced as a result of the specific types of model misspecification. Modification Indices demonstrated higher power rates when items were subject to under-factoring than when items were misspecified according to alternate-factoring. Wald Test statistics were also sensitive to the type of misspecification; power rates for this item-fit index were highest for items that were subject to alternate-factoring in comparison to those misspecified according to under-factoring.

7.2.Considerations for Future Research

The simulation design conditions in this study resulted from specific decisions made by the author and, though they were made with the intention of being generalizable to various test designs and sample populations, they reflect certain limitations that could be further examined. The number and proportion of estimation issues encountered in this dissertation must also be considered. Lastly, this section presents additional methods for the construction and validation of Q-matrices.

First, the number of dimensions, test lengths, and sample sizes in this dissertation were constrained by practical restrictions on time and computing resources as well as being informed by a review of previous literature. More extreme levels of these conditions are present in other research (see Baumgartner & Homburg, 1996, for example), which could potentially produce larger effects for those fit indices demonstrating sensitivity. Similarly, the range of the MDISC and MDIFF item parameters selected for the six item types reflect a subset of all possible discrimination and difficulty values. These values were selected to be representative of typical educational assessment conditions; factor loadings from the seminal study by Hu and Bentler (1998) can be shown to approximate MDISC values of 0.98 to 1.33 while the values employed by Beauducel and Wittman (2005) and Jackson (2007) correspond to MDISC values which range 0.44 to 1.61. As described in Chapter 3, the range of MDIFF values is also less extreme than those employed by Finch (2011) and Zhang and Stone (2008), which were approximately -2.0 to +2.0 and -5.0 to +5.0, respectively. Degree and type of model misspecification is another condition which could be further manipulated. Over-factoring, the inclusion of additional parameters in the estimating Q-matrix, was not included in this study and is not a condition to frequently appear in model misspecification research as this type of misspecification allows the number of potential models to explode very quickly, becoming again a matter of selection on the part of the researcher. Regarding the degree of misspecification, the majority of RMSEA values observed in the misspecified conditions ranged 0.02 to 0.15 which imply marginal-to-poor fit, therefore, the effects of more extreme misspecification could be explored. Fan

and Sivo (2005; 2007) discuss methods by which the degree of misspecification can be explicitly controlled other than direct manipulation of the Q-matrix.

Returning now to the issues encountered in the estimation of the misspecified models. Chen, Curran, Bollen, Kirby, and Paxton (2008) report a maximum of 29% of replicated models resulting in estimation issues; Fan, Thompson, and Wang (1999) reported that 3% of all replications of misspecified models resulted in estimation issues for sample sizes of 200 or greater; and Ximénez (2009) reported approximately 40% of misspecified model replications resulting in estimation issues. While these numbers are large, severely-misspecified models estimated for a sample size of 1000 with 3 weakly correlated factors, 12 high-discrimination / moderate-difficulty items which followed between-item multidimensionality in the current dissertation required a total of 107,725 replications to achieve 250 successful replications. For comparison, the estimation failure rate was 99.8% suggesting that this condition is essentially unestimable. In a preliminary study, models were estimated in this dissertation without requesting the output of factor scores, resulting in a percentage of rejected replications which were more aligned with the results of previous research. Upon requesting the output of factor scores, Mplus reported estimated inter-factor correlations that were greatly inflated, and greater than 1.0, which prevented subsequent estimation of factor scores. As a result of these findings, the degree of misspecification was lessened to produce fewer estimation failures. Further, the parameter recovery results demonstrate that the inter-factor correlation is highly sensitive to model misspecification and is poorly recovered when models are severely misspecified. These results suggest that previous research failed to output factor scores, therefore, sidestepping the resulting estimation issues. Designed to produce factor scores,

IRT software such as NOHARM (Fraser & McDonald, 1988) and Winsteps (Linacre, 2011) may provide robust estimation options which avoid related issues of estimation failure.

Decisions regarding both the initial selection of the simulation conditions and the subsequent revisions necessary due to observed estimation failures serve to limit the generalizability of the current study. In the first case, the specific levels of each simulation factor or condition represent but a sampling of all conditions possible. Further, these levels and conditions represent reasonable or feasible conditions under which the replications of the current study were expected to be successful. Consideration could be given to values beyond these ranges which may be considered unreasonable but still possible in the broader population. With regards to the revision of the simulation conditions as a result of the numerous estimation failures, the results of the current study for those specific conditions may be considered overly optimistic. Acknowledging the high rate of estimation failures, it may not be possible to generalize these results to other studies as the current study essentially describes results for conditions that cannot be successfully estimated. Having pursued successful estimation of such models and conditions, however, the current study describes conditions where successful estimation is likely not possible while also describing the performance of model- and item-fit indices should such models be successfully estimated.

The definition and construction of the Q-matrices in this dissertation was based on a very narrow sample of the vast population of Q-matrices that could be applied in both simulation and real data analysis. This study limited the Q-matrix both in the number of attributes and the number of associations permitted between those attributes and the

observed variables. Q-matrix definition is nearly unbounded, limited only by the imagination of researchers or constraints applied during the estimation process. Rupp, Templin, and Henson (2010) provide a good overview on the construction and interpretation of Q-matrices. Further, the Q-matrices employed in this study simply represent the associations between observed variables or test items and unobserved variables, such as latent constructs or abilities. This method of Q-matrix construction is said to be simple because it requires only consideration of the direct relationships between items and latent variables; the method is agnostic to strategies or methods employed by the subject or examinee in demonstrating the types of behavior necessary for success.

The *Attribute Hierarchy Method* (AHM; Leighton, Gierl, & Hunka, 2004) is an alternate approach to Q-matrix specification that has been proposed which takes into account the strategies necessary for successful performance and incorporates such dependencies in the final Q-matrix. In AHM, an initial assumption is made that the latent variables are considered to be hierarchically related or structured reflecting empirical and/or theoretical considerations. Next, a series of matrices (i.e., adjacency, reachability, incidence, and reduced incidence) are developed to represent performance profiles (see Tatsuoaka, 1983, 1995, 1996). The $k \times k$ *adjacency matrix* indicates the direct relationships posited between the latent variables; the $k \times k$ *reachability matrix* indicates the direct and indirect relationships between latent variables; the $k \times (2^k - 1)$ *incidence matrix* contains a single instance or item for each combination of attributes; and, lastly, the *reduced incidence matrix* retains only those columns from the incidence matrix which are logically permissible given the reachability matrix. This final matrix can be used

during the test development process to specify item types and demands to be assessed or transposed columns from the reduced incidence matrix could be used to construct a Q-matrix to be applied to an existing assessment. Following this alternate method of Q-matrix construction accounts for anticipated or hypothesized strategies and relationships between latent variables, not just the relationships between items and latent variables. This process is similar to the specification of attribute hierarchies in parametric diagnostic measurement models, which generally serve to reduce the complexity of the structural component of these models (see Rupp, Templin, & Henson, 2010; chapter 10).

Regardless of the method used in constructing the Q-matrix, Raykov and Marcoulides (2001) and Bechger, Verstralen, and Verhelst (2002) showed that it is possible for there to exist any number of equivalent models or Q-matrices. Therefore, researchers are encouraged to consider evidence beyond what is offered by model- and item-fit statistics in selecting a valid model estimated from real data. When estimating models applied to real data, the impact of the potential models on certain outcomes can provide additional evidence for the validity of the model. For example, the real data analysis example provided in this dissertation included an assessment designed to measure student-level math ability; therefore, outcomes estimated by this model should exhibit a reasonable degree of correspondence to other measures of math ability.

Further, if scores for the latent dimensions were estimated and interpreted, it could be expected that these scores would be associated with the results of similar measures (e.g., processing speed, general intelligence). Validity studies can be designed to evaluate the impact of the estimated model on such concurrent or criterion measures related to the initial assessment and/or latent dimensions. The validity of the selected

model, in comparison to any number of equivalent models, might also be examined through sensitivity to intervention activities applied to affect change in the latent dimensions. Should these dimensions truly represent skills or attributes then activities targeted at specifically affecting change should be seen to influence scores over the course of longitudinal observation. In addition to statistical measures of model- and item-fit, well-designed studies that seek to provide empirical evidence about the nomothetic span of the proposed latent dimensions provide important secondary evidence for model validity.

7.3. Conclusion

The theoretical cut points for each of the model- and item-fit indices were shown in this dissertation to produce inflated Type-I error rates; moreover, it was shown how suggested descriptive cut-off values for the χ^2/df and RMSEA statistic cannot be correctly used as cut-offs for hypothesis tests to control a nominal Type-I error rate. The results of this dissertation provide evidence for the use of cut points that account for the various sensitivities demonstrated by these fit indices. While a simulation study such as the one presented here is beyond the scope of most researchers wishing to evaluate model fit, a bootstrap approach as suggested by Tay and Drasgow (2012) may provide a reasonable option.

For the majority of conditions considered in this study, the model-fit indices demonstrated high power rates in correctly rejecting misspecified models. As a result of this, a two-stage approach to Q-matrix evaluation is presented wherein global misspecification is first evaluated via model-fit indices then local misspecification is explored by examining values of the item-fit indices. The results of such an approach

yields increased power rates for the item fit indices, suggesting that joint evaluation of model and item fit is likely to lead to appropriate revisions of misspecified Q-matrices. Increased consideration of item fit information and alternate approaches to the evaluation of model fit have been suggested by others such as Heene, Hilbert, Draxler, Ziegler, and Bühner (2011) and Saris, Satorra, and van der Veld (2009).

The amount of information presented in this dissertation can be distilled into two salient points: (1) fit index values under correctly specified models vary systematically across test design conditions, including the generally-expected effects of sample size and test length but also due to differences in item operating characteristics; and (2) though the power of item-fit indices to identify misspecified items is generally poor-to-moderate when design-appropriate empirical cut points are used, the power of model-fit indices is high and can be used to increase the likelihood of identifying Q-matrix misspecification when the two types of fit indices are jointly applied during model evaluation. Statistical power is demonstrated throughout this dissertation as the application of design-appropriate empirical cut points in rejecting misspecified models and items. However, the results of this dissertation also present power more generally, providing information on the most appropriate application of a wide variety of fit indices for the purpose of evaluating and refining Q-matrices, or measurement model structures, whether the models are estimated according to confirmatory factor analysis or multidimensional item response theory.

Appendix A

Q Matrices

Table A.1

Q-Matrix for 2 Latent Factors and 12 Items according to Within-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M	L	H	M
1	1	1		1	1		1	1		-2	-2	-2	1.4	0.9
2	1	1		1	1		0	1		-0.25	-0.75	-1	1.4	0.9
3	1	0		1	0		1	0		-0.1	-0.25	-0.75	1.6	1.1
4	1	0		1	0		1	0		1	-0.1	-0.5	1.6	1.1
5	0	1		0	1		0	1		1.13	0.1	-0.25	1.4	0.9
6	0	1		1	0		1	0		1.25	0.5	-0.1	1.4	0.9
7	1	0		0	1		0	1		1.38	1	0.1	1.4	0.9
8	1	0		1	0		1	0		1.5	1.2	0.25	1.4	0.9
9	0	1		0	1		0	1		1.63	1.4	0.5	1.6	1.1
10	0	1		0	1		0	1		1.75	1.6	0.75	1.6	1.1
11	1	1		1	1		1	0		1.88	1.8	1	1.4	0.9
12	1	1		1	1		1	1		2	2	2	1.4	0.9

Table A.2

Q-Matrix for 2 Latent Factors and 12 Items according to Between-Item Multidimensionality

Item	Q1	True	Q3	Q1	Mod.	Q3	Q1	Sev.	Q3	H	MDIFF	L	MDISC	
		Q2			Q2			Q2			M		H	M
1	1	0		1	0		1	0		-2	-2	-2	1.4	0.9
2	1	0		1	0		0	1		-0.25	-0.75	-1	1.4	0.9
3	1	0		1	0		1	0		-0.1	-0.25	-0.75	1.6	1.1
4	1	0		1	0		1	0		1	-0.1	-0.5	1.6	1.1
5	0	1		0	1		0	1		1.13	0.1	-0.25	1.4	0.9
6	0	1		1	0		1	0		1.25	0.5	-0.1	1.4	0.9
7	1	0		0	1		0	1		1.38	1	0.1	1.4	0.9
8	1	0		1	0		1	0		1.5	1.2	0.25	1.4	0.9
9	0	1		0	1		0	1		1.63	1.4	0.5	1.6	1.1
10	0	1		0	1		0	1		1.75	1.6	0.75	1.6	1.1
11	0	1		0	1		1	0		1.88	1.8	1	1.4	0.9
12	0	1		0	1		0	1		2	2	2	1.4	0.9

Table A.3

Q-Matrix for 2 Latent Factors and 24 Items according to Within-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M	L	H	M
1	1	1		1	1		1	1		-2	-2	-2	1.4	0.9
2	1	1		1	1		1	1		-0.75	-1.67	-1.67	1.4	0.9
3	1	1		1	1		0	1		-0.25	-0.75	-1.33	1.4	0.9
4	1	1		1	1		0	1		0.25	-0.667	-1	1.4	0.9
5	1	0		1	0		1	0		0.5	-0.25	-0.75	1.6	1.1
6	1	0		1	0		1	0		1	-0.179	-0.667	1.6	1.1
7	1	0		1	0		1	0		1.056	-0.107	-0.583	1.6	1.1
8	1	0		1	0		1	0		1.111	-0.036	-0.5	1.6	1.1
9	0	1		0	1		0	1		1.167	0.036	-0.25	1.4	0.9
10	0	1		0	1		0	1		1.222	0.107	-0.179	1.4	0.9
11	0	1		1	0		1	0		1.278	0.5	-0.107	1.4	0.9
12	0	1		1	0		1	0		1.333	0.583	-0.036	1.4	0.9
13	1	0		0	1		0	1		1.389	1	0.036	1.4	0.9
14	1	0		0	1		0	1		1.444	1	0.107	1.4	0.9
15	1	0		1	0		1	0		1.5	1	0.179	1.4	0.9
16	1	0		1	0		1	0		1.556	1	0.25	1.4	0.9
17	0	1		0	1		0	1		1.611	1	0.5	1.6	1.1
18	0	1		0	1		0	1		1.667	1.143	0.583	1.6	1.1
19	0	1		0	1		0	1		1.722	1.286	0.667	1.6	1.1
20	0	1		0	1		0	1		1.778	1.429	0.75	1.6	1.1
21	1	1		1	1		1	0		1.833	1.571	1	1.4	0.9
22	1	1		1	1		1	0		1.889	1.714	1.333	1.4	0.9
23	1	1		1	1		1	1		1.944	1.857	1.667	1.4	0.9
24	1	1		1	1		1	1		2	2	2	1.4	0.9

Table A.4

Q-Matrix for 2 Latent Factors and 24 Items according to Between-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M	L	H	M
1	1	0		1	0		1	0		-2	-2	-2	1.4	0.9
2	1	0		1	0		1	0		-0.75	-1.67	-1.67	1.4	0.9
3	1	0		1	0		0	1		-0.25	-0.75	-1.33	1.4	0.9
4	1	0		1	0		0	1		0.25	-0.667	-1	1.4	0.9
5	1	0		1	0		1	0		0.5	-0.25	-0.75	1.6	1.1
6	1	0		1	0		1	0		1	-0.179	-0.667	1.6	1.1
7	1	0		1	0		1	0		1.056	-0.107	-0.583	1.6	1.1
8	1	0		1	0		1	0		1.111	-0.036	-0.5	1.6	1.1
9	0	1		0	1		0	1		1.167	0.036	-0.25	1.4	0.9
10	0	1		0	1		0	1		1.222	0.107	-0.179	1.4	0.9
11	0	1		1	0		1	0		1.278	0.5	-0.107	1.4	0.9
12	0	1		1	0		1	0		1.333	0.583	-0.036	1.4	0.9
13	1	0		0	1		0	1		1.389	1	0.036	1.4	0.9
14	1	0		0	1		0	1		1.444	1	0.107	1.4	0.9
15	1	0		1	0		1	0		1.5	1	0.179	1.4	0.9
16	1	0		1	0		1	0		1.556	1	0.25	1.4	0.9
17	0	1		0	1		0	1		1.611	1	0.5	1.6	1.1
18	0	1		0	1		0	1		1.667	1.143	0.583	1.6	1.1
19	0	1		0	1		0	1		1.722	1.286	0.667	1.6	1.1
20	0	1		0	1		0	1		1.778	1.429	0.75	1.6	1.1
21	0	1		0	1		1	0		1.833	1.571	1	1.4	0.9
22	0	1		0	1		1	0		1.889	1.714	1.333	1.4	0.9
23	0	1		0	1		0	1		1.944	1.857	1.667	1.4	0.9
24	0	1		0	1		0	1		2	2	2	1.4	0.9

Table A.5

Q-Matrix for 2 Latent Factors and 36 Items according to Within-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M	L	H	M
1	1	1		1	1		1	1		-2	-2	-2	1.4	0.9
2	1	1		1	1		1	1		-0.75	-1.8	-1.8	1.4	0.9
3	1	1		1	1		1	1		-0.25	-1.6	-1.6	1.4	0.9
4	1	1		1	1		0	1		-0.2	-0.75	-1.4	1.4	0.9
5	1	1		1	1		0	1		-0.15	-0.7	-1.2	1.4	0.9
6	1	1		1	1		0	1		-0.1	-0.65	-1	1.4	0.9
7	1	0		1	0		1	0		-0.05	-0.25	-0.75	1.6	1.1
8	1	0		1	0		1	0		-0.01	-0.2	-0.7	1.6	1.1
9	1	0		1	0		1	0		0.5	-0.15	-0.65	1.6	1.1
10	1	0		1	0		1	0		1	-0.1	-0.6	1.6	1.1
11	1	0		1	0		1	0		1.038	-0.05	-0.55	1.6	1.1
12	1	0		1	0		1	0		1.077	-0.01	-0.5	1.6	1.1
13	0	1		0	1		0	1		1.115	0.01	-0.25	1.4	0.9
14	0	1		0	1		0	1		1.154	0.05	-0.2	1.4	0.9
15	0	1		0	1		0	1		1.192	0.1	-0.15	1.4	0.9
16	0	1		1	0		1	0		1.231	0.5	-0.1	1.4	0.9
17	0	1		1	0		1	0		1.269	0.625	-0.05	1.4	0.9
18	0	1		1	0		1	0		1.308	0.75	-0.01	1.4	0.9
19	1	0		0	1		0	1		1.346	1	0.01	1.4	0.9
20	1	0		0	1		0	1		1.385	1.059	0.05	1.4	0.9
21	1	0		0	1		0	1		1.423	1.118	0.1	1.4	0.9
22	1	0		1	0		1	0		1.462	1.176	0.15	1.4	0.9
23	1	0		1	0		1	0		1.5	1.235	0.2	1.4	0.9
24	1	0		1	0		1	0		1.538	1.294	0.25	1.4	0.9

Item	Q1	True	Q3	Q1	Mod.	Q3	Q1	Sev.	Q3	H	MDIFF	L	MDISC	M
		Q2			Q2			Q2			M		H	
25	0	1		0	1		0	1		1.577	1.353	0.5	1.6	1.1
26	0	1		0	1		0	1		1.615	1.412	0.55	1.6	1.1
27	0	1		0	1		0	1		1.654	1.471	0.6	1.6	1.1
28	0	1		0	1		0	1		1.692	1.529	0.65	1.6	1.1
29	0	1		0	1		0	1		1.731	1.588	0.7	1.6	1.1
30	0	1		0	1		0	1		1.769	1.647	0.75	1.6	1.1
31	1	1		1	1		1	0		1.808	1.706	1	1.4	0.9
32	1	1		1	1		1	0		1.846	1.765	1.2	1.4	0.9
33	1	1		1	1		1	0		1.885	1.824	1.4	1.4	0.9
34	1	1		1	1		1	1		1.923	1.882	1.6	1.4	0.9
35	1	1		1	1		1	1		1.962	1.941	1.8	1.4	0.9
36	1	1		1	1		1	1		2	2	2	1.4	0.9

Table A.6

Q-Matrix for 2 Latent Factors and 36 Items according to Between-Item Multidimensionality

Item	Q1	True	Q3	Q1	Mod.	Q3	Q1	Sev.	Q3	H	MDIFF	L	MDISC	M
		Q2			Q2			Q2			M		H	
1	1	0		1	0		1	0		-2	-2	-2	1.4	0.9
2	1	0		1	0		1	0		-0.75	-1.8	-1.8	1.4	0.9
3	1	0		1	0		1	0		-0.25	-1.6	-1.6	1.4	0.9
4	1	0		1	0		0	1		-0.2	-0.75	-1.4	1.4	0.9
5	1	0		1	0		0	1		-0.15	-0.7	-1.2	1.4	0.9
6	1	0		1	0		0	1		-0.1	-0.65	-1	1.4	0.9
7	1	0		1	0		1	0		-0.05	-0.25	-0.75	1.6	1.1
8	1	0		1	0		1	0		-0.01	-0.2	-0.7	1.6	1.1
9	1	0		1	0		1	0		0.5	-0.15	-0.65	1.6	1.1
10	1	0		1	0		1	0		1	-0.1	-0.6	1.6	1.1
11	1	0		1	0		1	0		1.038	-0.05	-0.55	1.6	1.1
12	1	0		1	0		1	0		1.077	-0.01	-0.5	1.6	1.1
13	0	1		0	1		0	1		1.115	0.01	-0.25	1.4	0.9
14	0	1		0	1		0	1		1.154	0.05	-0.2	1.4	0.9
15	0	1		0	1		0	1		1.192	0.1	-0.15	1.4	0.9
16	0	1		1	0		1	0		1.231	0.5	-0.1	1.4	0.9
17	0	1		1	0		1	0		1.269	0.625	-0.05	1.4	0.9
18	0	1		1	0		1	0		1.308	0.75	-0.01	1.4	0.9
19	1	0		0	1		0	1		1.346	1	0.01	1.4	0.9
20	1	0		0	1		0	1		1.385	1.059	0.05	1.4	0.9
21	1	0		0	1		0	1		1.423	1.118	0.1	1.4	0.9
22	1	0		1	0		1	0		1.462	1.176	0.15	1.4	0.9
23	1	0		1	0		1	0		1.5	1.235	0.2	1.4	0.9
24	1	0		1	0		1	0		1.538	1.294	0.25	1.4	0.9

Item	True			Mod.			Sev.			MDIFF			MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	H	M	L	H	M
25	0	1		0	1		0	1		1.577	1.353	0.5	1.6	1.1
26	0	1		0	1		0	1		1.615	1.412	0.55	1.6	1.1
27	0	1		0	1		0	1		1.654	1.471	0.6	1.6	1.1
28	0	1		0	1		0	1		1.692	1.529	0.65	1.6	1.1
29	0	1		0	1		0	1		1.731	1.588	0.7	1.6	1.1
30	0	1		0	1		0	1		1.769	1.647	0.75	1.6	1.1
31	0	1		0	1		1	0		1.808	1.706	1	1.4	0.9
32	0	1		0	1		1	0		1.846	1.765	1.2	1.4	0.9
33	0	1		0	1		1	0		1.885	1.824	1.4	1.4	0.9
34	0	1		0	1		0	1		1.923	1.882	1.6	1.4	0.9
35	0	1		0	1		0	1		1.962	1.941	1.8	1.4	0.9
36	0	1		0	1		0	1		2	2	2	1.4	0.9

Table A.7

Q-Matrix for 3 Latent Factors and 12 Items according to Within-Item Multidimensionality

Item	Q1	True		Q1	Mod.		Q1	Sev.		H	MDIFF		L	MDISC	
		Q2	Q3		Q2	Q3		Q2	Q3		M			H	M
1	1	1	0	1	1	0	1	0	0	-2	-2		-2	1.4	0.9
2	1	0	0	0	0	1	0	0	1	-0.25	-0.75		-1	1.5	1
3	0	1	0	0	1	0	0	1	0	-0.1	-0.25		-0.75	1.6	1.1
4	0	0	1	0	0	1	0	0	1	1	-0.1		-0.5	1.4	0.9
5	1	0	0	1	0	0	1	0	0	1.13	0.1		-0.25	1.5	1
6	1	0	1	1	0	1	1	0	1	1.25	0.5		-0.1	1.6	1.1
7	0	1	0	0	1	0	0	1	0	1.38	1		0.1	1.6	1.1
8	0	0	1	0	0	1	0	0	1	1.5	1.2		0.25	1.5	1
9	1	0	0	1	0	0	1	0	0	1.63	1.4		0.5	1.4	0.9
10	0	1	0	0	1	0	0	1	0	1.75	1.6		0.75	1.6	1.1
11	0	0	1	1	0	0	1	0	0	1.88	1.8		1	1.5	1
12	0	1	1	0	1	1	0	0	1	2	2		2	1.4	0.9

Table A.8

Q-Matrix for 3 Latent Factors and 12 Items according to Between-Item Multidimensionality

Item	Q1	True		Q1	Mod.		Q1	Sev.		H	MDIFF		L	MDISC	
		Q2	Q3		Q2	Q3		Q2	Q3		M			H	M
1	0	0	1	0	0	1	1	0	0	-2	-2		-2	1.4	0.9
2	1	0	0	0	0	1	0	0	1	-0.25	-0.75		-1	1.5	1
3	0	1	0	0	1	0	0	1	0	-0.1	-0.25		-0.75	1.6	1.1
4	0	0	1	0	0	1	0	0	1	1	-0.1		-0.5	1.4	0.9
5	1	0	0	1	0	0	1	0	0	1.13	0.1		-0.25	1.5	1
6	0	1	0	0	1	0	0	1	0	1.25	0.5		-0.1	1.6	1.1
7	0	1	0	0	1	0	0	1	0	1.38	1		0.1	1.6	1.1
8	0	0	1	0	0	1	0	0	1	1.5	1.2		0.25	1.5	1
9	1	0	0	1	0	0	1	0	0	1.63	1.4		0.5	1.4	0.9
10	0	1	0	0	1	0	0	1	0	1.75	1.6		0.75	1.6	1.1
11	0	0	1	1	0	0	1	0	0	1.88	1.8		1	1.5	1
12	1	0	0	1	0	0	0	0	1	2	2		2	1.4	0.9

Table A.9

Q-Matrix for 3 Latent Factors and 24 Items according to Within-Item Multidimensionality

Item	Q1	True		Q1	Mod.		Q1	Sev.		H	MDIFF		L	MDISC	
		Q2	Q3		Q2	Q3		Q2	Q3		M			H	M
1	1	1	0	1	1	0	1	0	0	-2	-2		-2	1.4	0.9
2	1	1	0	1	1	0	1	0	0	-0.75	-1.67		-1.67	1.4	0.9
3	1	0	0	0	0	1	0	0	1	-0.25	-0.75		-1.33	1.5	1
4	1	0	0	0	0	1	0	0	1	0.25	-0.667		-1	1.5	1
5	0	1	0	0	1	0	0	1	0	0.5	-0.25		-0.75	1.6	1.1
6	0	1	0	0	1	0	0	1	0	1	-0.179		-0.667	1.6	1.1
7	0	0	1	0	0	1	0	0	1	1.056	-0.107		-0.583	1.4	0.9
8	0	0	1	0	0	1	0	0	1	1.111	-0.036		-0.5	1.4	0.9
9	1	0	0	1	0	0	1	0	0	1.167	0.036		-0.25	1.5	1
10	1	0	0	1	0	0	1	0	0	1.222	0.107		-0.179	1.5	1
11	1	0	1	1	0	1	1	0	1	1.278	0.5		-0.107	1.6	1.1
12	1	0	1	1	0	1	1	0	1	1.333	0.583		-0.036	1.6	1.1
13	0	1	0	0	1	0	0	1	0	1.389	1		0.036	1.6	1.1
14	0	1	0	0	1	0	0	1	0	1.444	1		0.107	1.6	1.1
15	0	0	1	0	0	1	0	0	1	1.5	1		0.179	1.5	1
16	0	0	1	0	0	1	0	0	1	1.556	1		0.25	1.5	1
17	1	0	0	1	0	0	1	0	0	1.611	1		0.5	1.4	0.9
18	1	0	0	1	0	0	1	0	0	1.667	1.143		0.583	1.4	0.9
19	0	1	0	0	1	0	0	1	0	1.722	1.286		0.667	1.6	1.1
20	0	1	0	0	1	0	0	1	0	1.778	1.429		0.75	1.6	1.1
21	0	0	1	1	0	0	1	0	0	1.833	1.571		1	1.5	1
22	0	0	1	1	0	0	1	0	0	1.889	1.714		1.333	1.5	1
23	0	1	1	0	1	1	0	0	1	1.944	1.857		1.667	1.4	0.9
24	0	1	1	0	1	1	0	0	1	2	2		2	1.4	0.9

Table A.10

Q-Matrix for 3 Latent Factors and 24 Items according to Between-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M	L	H	M
1	0	0	1	0	0	1	1	0	0	-2	-2	-2	1.4	0.9
2	0	0	1	0	0	1	1	0	0	-0.75	-1.67	-1.67	1.4	0.9
3	1	0	0	0	0	1	0	0	1	-0.25	-0.75	-1.33	1.5	1
4	1	0	0	0	0	1	0	0	1	0.25	-0.667	-1	1.5	1
5	0	1	0	0	1	0	0	1	0	0.5	-0.25	-0.75	1.6	1.1
6	0	1	0	0	1	0	0	1	0	1	-0.179	-0.667	1.6	1.1
7	0	0	1	0	0	1	0	0	1	1.056	-0.107	-0.583	1.4	0.9
8	0	0	1	0	0	1	0	0	1	1.111	-0.036	-0.5	1.4	0.9
9	1	0	0	1	0	0	1	0	0	1.167	0.036	-0.25	1.5	1
10	1	0	0	1	0	0	1	0	0	1.222	0.107	-0.179	1.5	1
11	0	1	0	0	1	0	0	1	0	1.278	0.5	-0.107	1.6	1.1
12	0	1	0	0	1	0	0	1	0	1.333	0.583	-0.036	1.6	1.1
13	0	1	0	0	1	0	0	1	0	1.389	1	0.036	1.6	1.1
14	0	1	0	0	1	0	0	1	0	1.444	1	0.107	1.6	1.1
15	0	0	1	0	0	1	0	0	1	1.5	1	0.179	1.5	1
16	0	0	1	0	0	1	0	0	1	1.556	1	0.25	1.5	1
17	1	0	0	1	0	0	1	0	0	1.611	1	0.5	1.4	0.9
18	1	0	0	1	0	0	1	0	0	1.667	1.143	0.583	1.4	0.9
19	0	1	0	0	1	0	0	1	0	1.722	1.286	0.667	1.6	1.1
20	0	1	0	0	1	0	0	1	0	1.778	1.429	0.75	1.6	1.1
21	0	0	1	1	0	0	1	0	0	1.833	1.571	1	1.5	1
22	0	0	1	1	0	0	1	0	0	1.889	1.714	1.333	1.5	1
23	1	0	0	1	0	0	0	0	1	1.944	1.857	1.667	1.4	0.9
24	1	0	0	1	0	0	0	0	1	2	2	2	1.4	0.9

Table A.11

Q-Matrix for 3 Latent Factors and 36 Items according to Within-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		L	MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M			H	M
1	1	1	0	1	1	0	1	0	0	-2	-2		-2	1.4	0.9
2	1	1	0	1	1	0	1	0	0	-0.75	-1.8		-1.8	1.4	0.9
3	1	1	0	1	1	0	1	0	0	-0.25	-1.6		-1.6	1.4	0.9
4	1	0	0	0	0	1	0	0	1	-0.2	-0.75		-1.4	1.5	1
5	1	0	0	0	0	1	0	0	1	-0.15	-0.7		-1.2	1.5	1
6	1	0	0	0	0	1	0	0	1	-0.1	-0.65		-1	1.5	1
7	0	1	0	0	1	0	0	1	0	-0.05	-0.25		-0.75	1.6	1.1
8	0	1	0	0	1	0	0	1	0	-0.01	-0.2		-0.7	1.6	1.1
9	0	1	0	0	1	0	0	1	0	0.5	-0.15		-0.65	1.6	1.1
10	0	0	1	0	0	1	0	0	1	1	-0.1		-0.6	1.4	0.9
11	0	0	1	0	0	1	0	0	1	1.038	-0.05		-0.55	1.4	0.9
12	0	0	1	0	0	1	0	0	1	1.077	-0.01		-0.5	1.4	0.9
13	1	0	0	1	0	0	1	0	0	1.115	0.01		-0.25	1.5	1
14	1	0	0	1	0	0	1	0	0	1.154	0.05		-0.2	1.5	1
15	1	0	0	1	0	0	1	0	0	1.192	0.1		-0.15	1.5	1
16	1	0	1	1	0	1	1	0	1	1.231	0.5		-0.1	1.6	1.1
17	1	0	1	1	0	1	1	0	1	1.269	0.625		-0.05	1.6	1.1
18	1	0	1	1	0	1	1	0	1	1.308	0.75		-0.01	1.6	1.1
19	0	1	0	0	1	0	0	1	0	1.346	1		0.01	1.6	1.1
20	0	1	0	0	1	0	0	1	0	1.385	1.059		0.05	1.6	1.1
21	0	1	0	0	1	0	0	1	0	1.423	1.118		0.1	1.6	1.1
22	0	0	1	0	0	1	0	0	1	1.462	1.176		0.15	1.5	1
23	0	0	1	0	0	1	0	0	1	1.5	1.235		0.2	1.5	1
24	0	0	1	0	0	1	0	0	1	1.538	1.294		0.25	1.5	1

Item	Q1	True	Q3	Q1	Mod.	Q3	Q1	Sev.	Q3	H	MDIFF	L	MDISC	M
		Q2			Q2			Q2			M		H	
25	1	0	0	1	0	0	1	0	0	1.577	1.353	0.5	1.4	0.9
26	1	0	0	1	0	0	1	0	0	1.615	1.412	0.55	1.4	0.9
27	1	0	0	1		0	1	0	0	1.654	1.471	0.6	1.4	0.9
28	0	1	0	0	1	0	0	1	0	1.692	1.529	0.65	1.6	1.1
29	0	1	0	0	1	0	0	1	0	1.731	1.588	0.7	1.6	1.1
30	0	1	0	0	1	0	0	1	0	1.769	1.647	0.75	1.6	1.1
31	0	0	1	1	0	0	1	0	0	1.808	1.706	1	1.5	1
32	0	0	1	1	0	0	1	0	0	1.846	1.765	1.2	1.5	1
33	0	0	1	1	0	0	1	0	0	1.885	1.824	1.4	1.5	1
34	0	1	1	0	1	1	0	0	1	1.923	1.882	1.6	1.4	0.9
35	0	1	1	0	1	1	0	0	1	1.962	1.941	1.8	1.4	0.9
36	0	1	1	0	1	1	0	0	1	2	2	2	1.4	0.9

Table A.12

Q-Matrix for 3 Latent Factors and 36 Items according to Between-Item Multidimensionality

Item	True			Mod.			Sev.			H	MDIFF		L	MDISC	
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3		M			H	M
1	0	0	1	0	0	1	1	0	0	-2	-2		-2	1.4	0.9
2	0	0	1	0	0	1	1	0	0	-0.75	-1.8		-1.8	1.4	0.9
3	0	0	1	0	0	1	1	0	0	-0.25	-1.6		-1.6	1.4	0.9
4	1	0	0	0	0	1	0	0	1	-0.2	-0.75		-1.4	1.5	1
5	1	0	0	0	0	1	0	0	1	-0.15	-0.7		-1.2	1.5	1
6	1	0	0	0	0	1	0	0	1	-0.1	-0.65		-1	1.5	1
7	0	1	0	0	1	0	0	1	0	-0.05	-0.25		-0.75	1.6	1.1
8	0	1	0	0	1	0	0	1	0	-0.01	-0.2		-0.7	1.6	1.1
9	0	1	0	0	1	0	0	1	0	0.5	-0.15		-0.65	1.6	1.1
10	0	0	1	0	0	1	0	0	1	1	-0.1		-0.6	1.4	0.9
11	0	0	1	0	0	1	0	0	1	1.038	-0.05		-0.55	1.4	0.9
12	0	0	1	0	0	1	0	0	1	1.077	-0.01		-0.5	1.4	0.9
13	1	0	0	1	0	0	1	0	0	1.115	0.01		-0.25	1.5	1
14	1	0	0	1	0	0	1	0	0	1.154	0.05		-0.2	1.5	1
15	1	0	0	1	0		1	0	0	1.192	0.1		-0.15	1.5	1
16	0	1	0	0	1	0	0	1	0	1.231	0.5		-0.1	1.6	1.1
17	0	1	0	0	1	0	0	1	0	1.269	0.625		-0.05	1.6	1.1
18	0	1	0	0	1	0	0	1	0	1.308	0.75		-0.01	1.6	1.1
19	0	1	0	0	1	0	0	1	0	1.346	1		0.01	1.6	1.1
20	0	1	0	0	1	0	0	1	0	1.385	1.059		0.05	1.6	1.1
21	0	1	0	0	1	0	0	1	0	1.423	1.118		0.1	1.6	1.1
22	0	0	1	0	0	1	0	0	1	1.462	1.176		0.15	1.5	1
23	0	0	1	0	0	1	0	0	1	1.5	1.235		0.2	1.5	1
24	0	0	1	0	0	1	0	0	1	1.538	1.294		0.25	1.5	1

Item	Q1	True	Q3	Q1	Mod.	Q3	Q1	Sev.	Q3	H	MDIFF	L	MDISC	M
		Q2			Q2			Q2			M		H	
25	1	0	0	1	0	0	1	0	0	1.577	1.353	0.5	1.4	0.9
26	1	0	0	1	0	0	1	0	0	1.615	1.412	0.55	1.4	0.9
27	1	0	0	1	0	0	1	0	0	1.654	1.471	0.6	1.4	0.9
28	0	1	0	0	1	0	0	1	0	1.692	1.529	0.65	1.6	1.1
29	0	1	0	0	1	0	0	1	0	1.731	1.588	0.7	1.6	1.1
30	0	1	0	0	1	0	0	1	0	1.769	1.647	0.75	1.6	1.1
31	0	0	1	1	0	0	1	0	0	1.808	1.706	1	1.5	1
32	0	0	1	1	0	0	1	0	0	1.846	1.765	1.2	1.5	1
33	0	0	1	1	0	0	1	0	0	1.885	1.824	1.4	1.5	1
34	1	0	0	1	0	0	0	0	1	1.923	1.882	1.6	1.4	0.9
35	1	0	0	1	0	0	0	0	1	1.962	1.941	1.8	1.4	0.9
36	1	0	0	1	0	0	0	0	1	2	2	2	1.4	0.9

Appendix B

Key Descriptive Statistics Under True Model Estimation

Table B.1

Key Descriptive Statistics for the χ^2/df Model-Fit Index Under True Model Estimation

Test Length	Sample Size	Item Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
12	250	HH	0.655	0.964	1.086	1.060	1.173	1.307	1.404	1.661	3.391	0.180	1.570	7.425
12	250	HL	0.556	0.913	1.031	1.007	1.121	1.249	1.339	1.580	3.205	0.176	1.545	8.036
12	250	HM	0.644	0.948	1.057	1.033	1.137	1.261	1.347	1.572	3.346	0.167	1.769	10.020
12	250	MH	0.536	0.857	0.968	0.954	1.062	1.176	1.245	1.393	1.693	0.156	0.554	0.520
12	250	ML	0.453	0.873	0.995	0.981	1.102	1.223	1.301	1.466	2.092	0.174	0.518	0.635
12	250	MM	0.464	0.861	0.977	0.965	1.077	1.195	1.270	1.421	1.826	0.164	0.523	0.530
12	1000	HH	0.533	0.858	0.953	0.939	1.033	1.133	1.196	1.342	1.744	0.138	0.675	1.081
12	1000	HL	0.467	0.838	0.954	0.936	1.051	1.170	1.251	1.447	1.858	0.167	0.720	1.101
12	1000	HM	0.528	0.853	0.951	0.937	1.033	1.138	1.213	1.372	1.812	0.145	0.698	1.147
12	1000	MH	0.450	0.844	0.974	0.960	1.091	1.210	1.288	1.463	1.867	0.182	0.474	0.416
12	1000	ML	0.426	0.863	1.000	0.986	1.118	1.254	1.342	1.512	2.045	0.194	0.514	0.573
12	1000	MM	0.381	0.851	0.984	0.971	1.101	1.231	1.313	1.492	1.911	0.188	0.463	0.404
24	250	HH	0.876	0.987	1.026	1.020	1.057	1.097	1.124	1.191	1.388	0.056	0.802	1.479
24	250	HL	0.837	0.959	0.996	0.990	1.026	1.067	1.094	1.156	1.503	0.055	0.833	2.231
24	250	HM	0.848	0.969	1.005	0.999	1.035	1.073	1.101	1.161	1.475	0.053	0.960	2.875
24	250	MH	0.826	0.963	1.003	0.999	1.038	1.077	1.101	1.151	1.283	0.057	0.353	0.269
24	250	ML	0.797	0.962	1.006	1.003	1.046	1.089	1.117	1.172	1.305	0.064	0.328	0.274
24	250	MM	0.809	0.957	0.999	0.995	1.038	1.079	1.104	1.159	1.249	0.061	0.349	0.209
24	1000	HH	0.795	0.938	0.978	0.973	1.012	1.055	1.082	1.137	1.302	0.058	0.550	0.707
24	1000	HL	0.750	0.929	0.976	0.971	1.017	1.064	1.096	1.160	1.349	0.067	0.523	0.672
24	1000	HM	0.769	0.932	0.973	0.968	1.007	1.051	1.080	1.141	1.364	0.060	0.587	1.053
24	1000	MH	0.741	0.943	0.996	0.992	1.045	1.095	1.126	1.194	1.359	0.076	0.303	0.249

Test Length	Sample Size	Item Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
24	1000	ML	0.746	0.950	1.006	1.003	1.058	1.109	1.143	1.211	1.401	0.080	0.252	0.178
24	1000	MM	0.733	0.947	1.001	0.998	1.052	1.105	1.136	1.198	1.379	0.079	0.239	0.165
36	250	HH	0.916	0.987	1.008	1.004	1.025	1.047	1.060	1.090	1.220	0.029	0.740	1.235
36	250	HL	0.912	0.975	0.995	0.992	1.012	1.034	1.049	1.083	1.226	0.030	0.819	1.721
36	250	HM	0.929	0.985	1.004	1.001	1.019	1.040	1.054	1.084	1.232	0.028	1.038	3.070
36	250	MH	0.905	0.989	1.010	1.008	1.029	1.049	1.063	1.088	1.181	0.030	0.322	0.304
36	250	ML	0.892	0.984	1.008	1.006	1.030	1.052	1.067	1.095	1.173	0.034	0.290	0.218
36	250	MM	0.901	0.978	0.999	0.997	1.019	1.040	1.054	1.083	1.136	0.031	0.420	0.353
36	1000	HH	0.873	0.963	0.988	0.985	1.010	1.036	1.053	1.089	1.172	0.037	0.539	0.767
36	1000	HL	0.862	0.956	0.982	0.979	1.005	1.033	1.051	1.089	1.211	0.039	0.562	0.938
36	1000	HM	0.874	0.960	0.979	0.977	0.996	1.016	1.033	1.068	1.192	0.030	0.674	1.516
36	1000	MH	0.862	0.971	1.002	1.001	1.030	1.059	1.077	1.110	1.191	0.044	0.216	0.061
36	1000	ML	0.816	0.971	1.004	1.003	1.036	1.066	1.085	1.118	1.243	0.048	0.129	0.011
36	1000	MM	0.840	0.968	0.999	0.998	1.029	1.058	1.078	1.114	1.211	0.046	0.239	0.141

Table B.2

Key Descriptive Statistics for the RMSEA Model-Fit Index Under True Model Estimation

Test Length	Sample Size	Item Multi.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
12	250	B	0.000	0.000	0.013	0.009	0.024	0.033	0.038	0.049	0.098	0.014	0.800	-0.141
12	250	W	0.000	0.000	0.010	0.000	0.019	0.028	0.034	0.042	0.076	0.012	0.985	-0.136
12	1000	B	0.000	0.000	0.004	0.000	0.009	0.014	0.017	0.021	0.032	0.006	1.120	0.078
12	1000	W	0.000	0.000	0.004	0.000	0.008	0.013	0.016	0.021	0.032	0.006	1.387	0.864
24	250	B	0.000	0.000	0.008	0.007	0.015	0.019	0.022	0.027	0.045	0.008	0.560	-0.846
24	250	W	0.000	0.000	0.005	0.000	0.011	0.016	0.019	0.024	0.035	0.007	1.034	-0.108
24	1000	B	0.000	0.000	0.003	0.000	0.006	0.009	0.011	0.014	0.020	0.004	1.007	-0.224
24	1000	W	0.000	0.000	0.002	0.000	0.005	0.009	0.011	0.013	0.020	0.004	1.322	0.557
36	250	B	0.000	0.000	0.006	0.006	0.011	0.014	0.016	0.020	0.030	0.006	0.471	-0.958
36	250	W	0.000	0.000	0.004	0.000	0.008	0.012	0.014	0.017	0.027	0.005	1.052	-0.100
36	1000	B	0.000	0.000	0.002	0.000	0.005	0.007	0.009	0.010	0.015	0.003	1.005	-0.252
36	1000	W	0.000	0.000	0.002	0.000	0.004	0.007	0.008	0.010	0.016	0.003	1.374	0.704

Table B.3

Key Descriptive Statistics for the GDDM Model-Fit Index Under True Model Estimation

Test Length	Sample Size	Item Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
12	250	HH	1.569	1.844	1.915	1.906	1.972	2.049	2.106	2.237	2.931	0.108	0.909	2.941
12	250	HL	1.249	1.676	1.741	1.738	1.797	1.867	1.922	2.048	2.401	0.107	0.511	1.876
12	250	HM	1.467	1.743	1.804	1.801	1.857	1.916	1.963	2.091	2.498	0.096	0.628	2.059
12	250	MH	1.474	1.779	1.867	1.852	1.942	2.037	2.102	2.230	3.038	0.130	0.754	1.674
12	250	ML	1.115	1.560	1.645	1.633	1.717	1.813	1.878	2.026	2.428	0.130	0.627	1.342
12	250	MM	1.374	1.643	1.716	1.708	1.778	1.856	1.907	2.033	3.180	0.113	0.933	5.035
12	1000	HH	6.647	7.347	7.496	7.496	7.644	7.775	7.859	8.032	8.671	0.224	0.049	0.311
12	1000	HL	5.901	6.583	6.760	6.815	6.945	7.036	7.088	7.188	7.530	0.247	-0.554	-0.401
12	1000	HM	6.297	6.985	7.138	7.171	7.299	7.406	7.471	7.602	7.970	0.229	-0.382	-0.130
12	1000	MH	6.231	6.954	7.120	7.117	7.288	7.440	7.534	7.745	8.476	0.256	0.080	0.369
12	1000	ML	5.371	6.084	6.238	6.265	6.404	6.517	6.591	6.764	7.371	0.240	-0.248	0.037
12	1000	MM	5.796	6.515	6.661	6.679	6.822	6.945	7.021	7.162	7.545	0.235	-0.284	-0.017
24	250	HH	0.415	0.493	0.519	0.516	0.540	0.565	0.581	0.613	0.694	0.036	0.475	0.436
24	250	HL	0.352	0.448	0.479	0.475	0.505	0.534	0.555	0.595	0.709	0.042	0.563	0.603
24	250	HM	0.381	0.482	0.513	0.509	0.540	0.571	0.592	0.630	0.749	0.044	0.519	0.458
24	250	MH	0.395	0.504	0.534	0.531	0.561	0.590	0.609	0.646	1.135	0.043	0.651	3.513
24	250	ML	0.332	0.446	0.481	0.477	0.512	0.546	0.567	0.612	0.720	0.049	0.448	0.297
24	250	MM	0.359	0.483	0.520	0.516	0.553	0.590	0.614	0.662	0.775	0.053	0.458	0.262
24	1000	HH	1.690	1.830	1.881	1.873	1.923	1.974	2.010	2.071	2.286	0.070	0.580	0.459
24	1000	HL	1.400	1.605	1.650	1.653	1.694	1.738	1.769	1.829	2.020	0.072	0.095	0.499
24	1000	HM	1.573	1.738	1.788	1.782	1.831	1.885	1.920	1.991	2.188	0.074	0.510	0.733
24	1000	MH	1.562	1.803	1.860	1.855	1.911	1.967	2.004	2.082	2.316	0.083	0.454	0.573
24	1000	ML	1.314	1.505	1.560	1.556	1.611	1.667	1.701	1.773	1.959	0.082	0.347	0.316
24	1000	MM	1.447	1.663	1.729	1.722	1.786	1.852	1.896	1.987	2.152	0.094	0.499	0.428
36	250	HH	0.210	0.259	0.274	0.273	0.288	0.303	0.313	0.330	0.383	0.022	0.341	0.163

Test Length	Sample Size	Item Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
36	250	HL	0.176	0.238	0.255	0.254	0.271	0.287	0.298	0.318	0.368	0.024	0.351	0.183
36	250	HM	0.196	0.244	0.260	0.259	0.274	0.289	0.299	0.320	0.366	0.023	0.374	0.251
36	250	MH	0.211	0.275	0.293	0.292	0.309	0.327	0.337	0.357	0.409	0.026	0.293	0.159
36	250	ML	0.185	0.249	0.269	0.268	0.287	0.306	0.317	0.340	0.391	0.028	0.310	0.134
36	250	MM	0.193	0.258	0.276	0.275	0.294	0.311	0.323	0.345	0.394	0.027	0.284	0.126
36	1000	HH	0.750	0.851	0.882	0.878	0.909	0.939	0.958	0.998	1.067	0.042	0.518	0.293
36	1000	HL	0.638	0.752	0.781	0.780	0.809	0.838	0.857	0.893	0.986	0.044	0.193	0.255
36	1000	HM	0.722	0.802	0.829	0.826	0.852	0.878	0.894	0.928	1.022	0.038	0.436	0.353
36	1000	MH	0.747	0.862	0.896	0.893	0.926	0.958	0.979	1.020	1.115	0.048	0.360	0.215
36	1000	ML	0.591	0.729	0.764	0.761	0.795	0.830	0.851	0.893	1.022	0.050	0.355	0.190
36	1000	MM	0.666	0.794	0.827	0.824	0.857	0.891	0.911	0.951	1.027	0.048	0.358	0.152

Table B.4

Key Descriptive Statistics for the $S\text{-}\chi^2/df$ Item-Fit Index Under True Model Estimation

Test Length	Sample Size	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
12	250	H	0.025	5.384	8.552	7.580	10.406	13.706	16.155	23.390	1049.197	8.239	35.224	2586.406
12	250	L	0.002	3.148	5.512	4.901	7.205	9.814	11.664	15.714	35.360	3.234	1.204	2.310
12	250	M	0.009	3.973	6.391	5.801	8.179	10.842	12.714	16.767	191.404	3.388	1.925	41.936
12	1000	H	0.367	11.074	17.706	15.943	22.291	29.610	35.687	50.105	249.584	9.514	1.417	4.578
12	1000	L	0.029	4.935	7.697	7.096	9.787	12.764	14.877	19.383	53.894	3.842	1.030	1.874
12	1000	M	0.173	7.378	11.100	10.350	13.977	18.002	20.811	26.966	52.949	5.239	0.944	1.532
24	250	H	0.122	9.465	13.368	12.669	16.387	20.296	22.971	29.101	883.094	6.664	17.832	1298.712
24	250	L	0.028	7.471	11.143	10.525	14.141	17.911	20.404	25.641	58.741	5.080	0.759	0.918
24	250	M	0.064	8.197	11.877	11.271	14.873	18.639	21.152	26.457	51.549	5.096	0.751	0.922
24	1000	H	1.531	17.537	23.830	22.685	28.773	35.362	40.258	51.394	129.345	9.003	0.923	1.700
24	1000	L	0.503	11.429	15.559	14.929	19.002	23.237	26.075	32.005	61.481	5.820	0.684	0.863
24	1000	M	0.994	13.986	18.535	17.895	22.387	27.032	30.128	36.715	68.149	6.490	0.646	0.909
36	250	H	0.020	12.263	16.995	16.327	20.919	25.599	28.692	35.262	877.334	7.003	5.923	393.160
36	250	L	0.032	10.421	15.055	14.472	19.020	23.570	26.526	32.611	60.133	6.428	0.579	0.465
36	250	M	0.176	11.101	15.765	15.157	19.717	24.320	27.304	33.504	60.870	6.449	0.595	0.489
36	1000	H	2.639	24.299	31.093	30.049	36.689	43.963	49.011	59.974	119.123	9.969	0.756	1.461
36	1000	L	0.915	17.416	22.588	21.926	27.037	32.292	35.711	42.929	84.508	7.383	0.566	0.690
36	1000	M	1.421	20.296	25.735	25.091	30.477	35.960	39.619	47.702	93.720	7.914	0.554	0.877

Table B.5

Key Descriptive Statistics for Modification Index 1 Under True Model Estimation

Corr.	Sample Size	Dim.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
H	250	2	0.000	0.097	0.994	0.435	1.287	2.680	3.860	6.945	32.864	1.469	3.243	17.640
H	250	3	0.000	0.077	0.782	0.342	1.007	2.112	3.045	5.491	23.134	1.155	3.164	15.799
H	1000	2	0.000	0.139	1.399	0.615	1.800	3.758	5.453	9.846	38.154	2.065	3.156	15.508
H	1000	3	0.000	0.113	1.113	0.494	1.447	3.002	4.308	7.660	33.481	1.619	3.065	14.666
L	250	2	0.000	0.178	1.800	0.795	2.336	4.844	6.972	12.488	71.149	2.638	3.178	16.843
L	250	3	0.000	0.136	1.364	0.606	1.775	3.688	5.299	9.310	35.164	1.982	3.042	14.418
L	1000	2	0.000	0.236	2.403	1.053	3.107	6.466	9.365	16.662	74.634	3.531	3.123	15.415
L	1000	3	0.000	0.179	1.805	0.805	2.348	4.865	6.988	12.369	45.988	2.616	3.004	13.746
M	250	2	0.000	0.145	1.459	0.644	1.891	3.938	5.664	10.108	40.309	2.132	3.097	15.018
M	250	3	0.000	0.114	1.147	0.510	1.494	3.091	4.439	7.868	35.917	1.668	3.083	15.201
M	1000	2	0.000	0.198	2.026	0.893	2.627	5.453	7.875	14.087	58.378	2.978	3.165	16.075
M	1000	3	0.000	0.158	1.570	0.703	2.050	4.233	6.059	10.722	42.264	2.268	3.015	14.131

Table B.6

Key Descriptive Statistics for Modification Index 2 Under True Model Estimation

Corr.	Sample Size	Dim.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
H	250	2	0.000	0.103	1.097	0.465	1.383	2.943	4.308	7.975	61.844	1.689	3.831	31.155
H	250	3	0.000	0.080	0.824	0.358	1.061	2.222	3.210	5.778	25.763	1.222	3.251	17.434
H	1000	2	0.000	0.143	1.433	0.634	1.852	3.865	5.581	9.967	36.341	2.101	3.100	14.783
H	1000	3	0.000	0.117	1.164	0.514	1.502	3.136	4.521	8.049	31.514	1.704	3.113	15.091
L	250	2	0.000	0.175	1.776	0.781	2.292	4.788	6.909	12.296	59.575	2.613	3.220	17.410
L	250	3	0.000	0.148	1.515	0.666	1.956	4.079	5.870	10.579	46.767	2.240	3.263	17.568
L	1000	2	0.000	0.233	2.379	1.048	3.096	6.441	9.236	16.365	63.426	3.469	3.061	14.629
L	1000	3	0.000	0.197	1.991	0.879	2.579	5.369	7.712	13.816	61.323	2.917	3.135	15.718
M	250	2	0.000	0.146	1.477	0.655	1.917	3.988	5.728	10.168	60.358	2.158	3.199	18.143
M	250	3	0.000	0.122	1.246	0.550	1.615	3.359	4.831	8.646	43.595	1.829	3.170	16.305
M	1000	2	0.000	0.196	2.017	0.886	2.612	5.431	7.846	14.043	50.223	2.964	3.127	15.358
M	1000	3	0.000	0.169	1.691	0.749	2.198	4.552	6.562	11.692	51.361	2.464	3.053	14.402

Table B.7

Key Descriptive Statistics for Modification Index 3 Under True Model Estimation

Corr.	Sample Size	Test Length	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
H	250	12	0.000	0.089	0.872	0.391	1.141	2.336	3.355	5.909	23.883	1.263	3.166	16.562
H	250	24	0.000	0.083	0.837	0.372	1.091	2.258	3.223	5.755	18.633	1.218	3.085	14.918
H	250	36	0.000	0.073	0.759	0.329	0.979	2.056	2.965	5.288	16.309	1.119	3.105	14.827
H	1000	12	0.000	0.103	0.981	0.443	1.285	2.643	3.779	6.626	20.199	1.403	2.942	13.220
H	1000	24	0.000	0.116	1.133	0.508	1.481	3.052	4.374	7.709	26.582	1.634	2.990	13.606
H	1000	36	0.000	0.119	1.178	0.522	1.529	3.174	4.573	8.125	42.498	1.713	3.080	15.352
L	250	12	0.000	0.118	1.175	0.520	1.514	3.162	4.575	8.162	23.693	1.719	3.042	13.867
L	250	24	0.000	0.135	1.359	0.600	1.767	3.667	5.291	9.387	34.637	1.974	3.000	13.722
L	250	36	0.000	0.138	1.397	0.618	1.812	3.749	5.396	9.698	38.330	2.057	3.226	17.015
L	1000	12	0.000	0.134	1.299	0.588	1.709	3.498	4.975	8.722	25.510	1.851	2.905	12.722
L	1000	24	0.000	0.178	1.741	0.786	2.282	4.706	6.722	11.597	39.082	2.489	2.947	13.404
L	1000	36	0.000	0.195	1.970	0.883	2.573	5.316	7.599	13.456	47.901	2.850	3.027	14.191
M	250	12	0.000	0.107	1.043	0.474	1.365	2.781	4.003	7.090	28.023	1.503	3.098	15.560
M	250	24	0.000	0.117	1.180	0.525	1.541	3.172	4.565	8.111	24.494	1.711	3.027	13.971
M	250	36	0.000	0.117	1.168	0.520	1.519	3.148	4.515	8.036	35.831	1.697	3.078	15.026
M	1000	12	0.000	0.122	1.177	0.533	1.549	3.153	4.544	7.988	21.466	1.682	2.925	12.744
M	1000	24	0.000	0.157	1.562	0.700	2.048	4.226	6.025	10.492	37.916	2.238	2.940	13.285
M	1000	36	0.000	0.173	1.726	0.771	2.258	4.647	6.653	11.795	36.322	2.489	2.977	13.445

Table B.8
Key Descriptive Statistics for Wald Test 1 Under True Model Estimation

Item Dim.	Sample Size	Test Length	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
B	250	12	-1.057	14.132	11.409	10.636	8.075	6.199	5.171	3.382	42.443	4.596	0.783	0.644
B	250	24	0.262	17.379	14.019	13.032	9.958	7.851	6.694	4.487	52.599	5.506	0.784	0.642
B	250	36	-0.049	18.414	14.819	13.751	10.505	8.332	7.181	5.162	55.000	5.769	0.783	0.530
B	1000	12	3.336	26.937	21.941	20.449	16.236	13.355	11.865	9.593	55.259	7.471	0.611	-0.305
B	1000	24	4.106	33.040	26.625	24.892	19.720	16.526	14.712	11.729	64.095	8.817	0.528	-0.455
B	1000	36	5.418	34.999	28.135	26.190	20.787	17.378	15.645	12.954	68.048	9.347	0.565	-0.477
W	250	12	-1.676	4.877	3.671	3.479	2.280	1.330	0.779	-0.205	14.597	1.963	0.519	0.320
W	250	24	-2.343	5.853	4.548	4.344	3.044	2.015	1.443	0.421	17.803	2.113	0.513	0.342
W	250	36	-2.570	6.260	4.931	4.713	3.384	2.335	1.771	0.798	20.202	2.162	0.540	0.330
W	1000	12	-1.223	8.878	6.988	6.674	4.735	3.244	2.497	1.375	20.733	3.080	0.545	0.097
W	1000	24	-0.408	10.690	8.622	8.352	6.183	4.561	3.768	2.438	23.225	3.305	0.489	0.067
W	1000	36	-0.207	11.462	9.368	9.067	6.890	5.255	4.463	3.193	23.686	3.352	0.502	0.031

Table B.9
Key Descriptive Statistics for Wald Test 2 Under True Model Estimation

Item Dim.	Sample Size	Test Length	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
B	250	HH	1.615	18.908	15.854	15.148	11.979	9.602	8.368	6.330	131.289	5.493	1.091	3.680
B	250	HM	2.012	20.610	17.481	17.167	13.888	11.123	9.633	7.320	75.702	5.165	0.573	1.108
B	250	HL	1.722	23.453	20.278	20.039	16.910	13.971	11.914	7.864	66.262	5.269	0.286	0.740
B	250	MH	0.451	11.882	10.100	9.764	7.954	6.541	5.777	4.456	32.298	3.026	0.709	0.982
B	250	MM	0.377	12.958	11.115	10.899	9.032	7.485	6.607	5.095	32.596	2.988	0.491	0.579
B	250	ML	1.145	14.290	12.401	12.211	10.305	8.690	7.753	6.017	35.974	3.042	0.402	0.490
B	1000	HH	8.482	32.358	28.433	28.067	24.138	20.826	18.957	15.889	60.631	6.095	0.313	-0.008
B	1000	HM	9.715	37.334	32.570	32.934	27.801	23.229	20.873	17.280	63.758	6.845	-0.073	-0.277
B	1000	HL	10.417	43.055	38.431	38.926	34.549	29.375	25.456	18.422	66.212	7.036	-0.488	0.640
B	1000	MH	6.440	21.045	18.794	18.491	16.235	14.331	13.246	11.415	36.856	3.664	0.428	0.163
B	1000	MM	7.739	23.573	20.979	20.967	18.323	15.806	14.493	12.320	38.620	3.926	0.081	-0.125
B	1000	ML	8.107	26.598	23.674	23.829	20.909	18.244	16.583	13.489	40.101	4.141	-0.181	-0.098
W	250	HH	-4.316	5.626	3.988	3.959	2.242	0.806	0.014	-1.147	17.444	2.453	0.178	-0.114
W	250	HM	-2.426	5.566	4.112	3.962	2.522	1.317	0.637	-0.376	16.863	2.242	0.363	0.083
W	250	HL	-1.458	6.061	4.690	4.516	3.146	2.074	1.500	0.517	16.338	2.138	0.448	0.180
W	250	MH	-2.407	4.646	3.453	3.321	2.138	1.178	0.623	-0.320	13.168	1.852	0.377	0.122
W	250	MM	-2.136	4.586	3.458	3.362	2.219	1.288	0.754	-0.171	14.417	1.751	0.338	0.161
W	250	ML	-1.917	4.916	3.774	3.689	2.533	1.591	1.057	0.089	13.074	1.755	0.302	0.112
W	1000	HH	-1.476	10.803	8.454	8.378	5.971	3.918	2.616	0.794	22.094	3.565	0.150	-0.219
W	1000	HM	-0.628	10.810	8.743	8.565	6.454	4.796	3.908	2.497	21.747	3.148	0.339	-0.084
W	1000	HL	0.284	11.701	9.435	9.248	7.025	5.403	4.554	3.186	21.892	3.178	0.232	-0.475
W	1000	MH	-1.279	8.667	6.857	6.750	4.924	3.531	2.788	1.489	17.420	2.615	0.200	-0.349
W	1000	MM	-1.177	8.721	6.972	6.945	5.098	3.767	3.076	1.886	18.284	2.481	0.151	-0.382
W	1000	ML	-0.840	9.223	7.392	7.384	5.470	4.127	3.445	2.230	17.731	2.503	0.089	-0.523

Table B.10
Key Descriptive Statistics for Wald Test 3 Under True Model Estimation

Item Dim.	Sample Size	Test Length	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
B	250	12	-1.152	11.484	9.099	8.619	6.324	4.353	3.132	1.492	42.415	4.003	0.611	0.545
B	250	24	-0.912	14.566	11.687	10.872	8.293	6.345	5.147	2.765	46.684	4.766	0.735	0.692
B	250	36	-0.600	15.429	12.446	11.551	8.910	6.943	5.770	3.413	58.543	4.949	0.802	0.953
B	1000	12	1.831	21.563	17.524	16.600	13.095	10.218	8.397	4.617	46.658	6.328	0.476	0.006
B	1000	24	3.454	27.478	22.261	20.630	16.909	14.109	12.470	8.729	52.629	7.269	0.549	-0.325
B	1000	36	4.264	28.963	23.594	21.797	18.006	15.265	13.706	10.661	55.421	7.464	0.607	-0.296
W	250	12	-1.693	5.054	3.794	3.590	2.367	1.368	0.793	-0.236	14.298	2.024	0.485	0.215
W	250	24	-2.281	6.134	4.756	4.521	3.159	2.103	1.512	0.472	15.542	2.222	0.506	0.152
W	250	36	-2.586	6.545	5.119	4.881	3.471	2.390	1.789	0.787	23.057	2.272	0.500	0.153
W	1000	12	-0.703	9.258	7.338	6.980	5.015	3.525	2.757	1.552	20.811	3.193	0.586	0.132
W	1000	24	0.388	11.190	9.100	8.637	6.495	4.916	4.131	2.864	24.470	3.538	0.627	0.085
W	1000	36	0.352	11.939	9.768	9.271	7.084	5.477	4.684	3.388	25.030	3.627	0.619	0.032

Appendix C

Investigation into Convergence and Replication Issues

When determining the total number of replications, precision and stability of the resulting estimates as well as feasibility and computing time must be considered. An analysis of convergence issues showed that the most stable model – a 2-factor model with low inter-factor correlation, 36 high-discrimination/low-difficulty (HL) items, one-half of which demonstrated within-item multidimensionality, and 1,000 examinees – produced no estimation failures while the success rate for the least stable model – a 3-factor of high inter-factor correlation, 12 moderately-discriminating/high-difficulty items of within-item multidimensionality, and 250 examinees – was less than 25%, indicating a number of estimation failures. An analysis of the standard errors of the key distributional characteristics from the empirical sampling distributions (i.e., mean, median, 90th percentile, 95th percentile) of the outcome statistics of interest (e.g., $S-\chi^2$ statistic, GDDM) indicated reasonable stable standard errors when about 100 to 200 replications were used.

The computing time required to estimate 1000 replications of the least stable model was approximately 20 minutes; the total time to estimate 1000 replications of all 864 experimental cells, therefore, would be approximately 18,000 minutes or 300 hours. The required computing time to achieve 100, 200, or 250 successful replications can be interpolated from these results as approximately 44, 90, and 110 hours, respectively. Since the computing time for 1000 replications is excessive given the desired time frame for completion of this dissertation, 250 replications of the misspecified model were chosen as a reasonable compromise between the statistical desideratum for reasonable

precision and stability of sampling distribution estimates and practical feasibility. The full 1000 replications are employed in the estimation of the true models.

C.1. Non-Convergent and Heywood Cases

Examining the 1000 replications of the model estimated according to the most stable conditions (correct or true model specification; 2 factors, low inter-factor correlation, 36 highly-discriminating/moderately-difficult items of within-item multidimensionality, and 1,000 examinees) evidenced no estimation issues; no replacement replications were required. The least stable boundary condition was identified as the estimation of 3 factors of high inter-factor correlation, 12 moderately-discriminating/low-difficulty items of within-item multidimensionality, 250 examinees. Suggested by previous research (Jackson, 2007), moderately misspecified models were estimated in anticipation that they would result in the greatest proportion of estimation failures. To achieve 1000 successful replications, a total of 4234 replications were required, a 23.6% success rate. This success rate for replications is smaller than that seen in previous research (Fan, Thompson, & Wang, 1999; Ximénez, 2009), suggesting that the degree of misspecification and other simulation conditions differ substantially.

C.2. Determining the Optimal Number of Replications

A study was conducted to determine the number of replications necessary to accurately describe the performance of the model- and item-fit statistics considered in the full study. Specifically, the objective was to find a cut-off for the number of replications beyond which increases in stability of the distributional characteristics (i.e., changes in estimated standard errors) would be practically negligible. Two experimental cells, one

representing a case where one would expect relatively stable model estimations and one where one would expect relatively unstable model estimations, were first identified. For each of these two cells 1,000 replications of the data-generation and model estimation process were computed. Then, using a bootstrapping method, 100 random samples of varying numbers of replications were drawn with replacement from the set of 1,000 replications. Specifically, 100 random draws of sizes 10, 50, 100, 200, 250, 500, and 1000 were made from the 1,000 replications and the mean and standard deviation (i.e., standard error) of key distributional indicators (i.e., mean, median, 90th percentile, 95th percentile, skewness, kurtosis) were then calculated across replication sets for each fit index.

An accurate assessment of the performance of model- and item- fit statistics in this Monte Carlo simulation study is also affected by missing information due to non-convergence of estimation and improper parameter estimates (e.g., Heywood cases). Notably, the omission or elimination of these replications would result in an unbalanced simulation design because different experimental cells would have different numbers of replications associated with them. To avoid this scenario, Heywood cases and non-converged models will be replaced with additional replications to ensure a balanced design as in previous studies (Fan, Thompson, & Wang, 1999; Jackson, 2007; Ximénez, 2009). The number of additional replications necessary to achieve the number suggested by the previous analysis for each of the two representative simulation conditions described above will be computed and used to inform the number of replications in the full study alongside the estimates of the distributional characteristics noted earlier.

C.3. Replications and the Two Factor Model

Representing the most stable estimation conditions, a 2-factor model was estimated for which the intra-factor correlation was specified as low and responses to 36 high-discrimination/low-difficulty (HL) items, one-half of which demonstrated within-item multidimensionality, were simulated for a sample size of 1,000 examinees.

The mean and variance of the distributional indicators and key indicators according to partition are presented for each of the model fit statistics in Table C.1. Figure C.1 depicts the model fit indices graphically with different plotting symbols representing the various values of the distributional and key indicators: empty circles represent mean values, stars represent median values, shaded squares represent the 90th percentile, shaded circles represent the 95th percentile, and shaded triangles represent the 99th percentile. For each model-fit index, point estimates and dispersion of the mean, median, and standard deviation are near-constant across replication sets, with the exception of the median RMSEA which decreases with number of replications due to the increasing proportion of replications where RMSEA is zero. The key indicators (90th, 95th, and 99th percentiles) generally appear to be stable at 100 replications and greater. Point estimates for each of the key indicators are differentiated at the hundredths decimal place for the χ^2/df and the thousandths decimal place for the RMSEA; key indicators of the GDDM are not well-differentiated even at the thousandths decimal place due to the fact that values of the GDDM are extremely small. Variability of the key indicators is typically less than 0.001, decreasing over replication sets with the largest decreases occurring between 10, 50, and 100 replications.

Table C.1

2-Factor Model: Distributional and Key Indicators for Model Fit Indices Across Partition Sets

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
χ^2/df	10	(1) mean	0.983	0.980	0.031	0.246	0.061	1.019	1.028	1.035
		(2) var	0.000	0.000	0.000	0.220	1.544	0.000	0.000	0.000
	50	(1) mean	0.982	0.979	0.033	0.408	0.547	1.022	1.036	1.062
		(2) var	0.000	0.000	0.000	0.131	1.034	0.000	0.000	0.000
	100	(1) mean	0.982	0.979	0.033	0.475	0.609	1.024	1.037	1.066
		(2) var	0.000	0.000	0.000	0.068	0.473	0.000	0.000	0.000
	200	(1) mean	0.982	0.979	0.033	0.516	0.666	1.024	1.038	1.071
		(2) var	0.000	0.000	0.000	0.048	0.396	0.000	0.000	0.000
	250	(1) mean	0.982	0.979	0.033	0.523	0.701	1.024	1.039	1.073
		(2) var	0.000	0.000	0.000	0.035	0.298	0.000	0.000	0.000
	500	(1) mean	0.982	0.979	0.033	0.514	0.700	1.024	1.038	1.075
		(2) var	0.000	0.000	0.000	0.020	0.145	0.000	0.000	0.000
	1000	(1) mean	0.982	0.979	0.033	0.539	0.742	1.025	1.038	1.078
		(2) var	0.000	0.000	0.000	0.008	0.074	0.000	0.000	0.000
RMSEA	10	(1) mean	0.001	0.000	0.002	1.292	2.863	0.004	0.005	0.006
		(2) var	0.000	0.000	0.000	0.338	14.898	0.000	0.000	0.000
	50	(1) mean	0.001	0.000	0.002	1.796	3.011	0.005	0.006	0.008
		(2) var	0.000	0.000	0.000	0.215	7.360	0.000	0.000	0.000
	100	(1) mean	0.001	0.000	0.002	1.842	2.763	0.005	0.006	0.008
		(2) var	0.000	0.000	0.000	0.099	2.175	0.000	0.000	0.000
	200	(1) mean	0.001	0.000	0.002	1.853	2.683	0.005	0.006	0.008
		(2) var	0.000	0.000	0.000	0.062	1.468	0.000	0.000	0.000
	250	(1) mean	0.001	0.000	0.002	1.860	2.675	0.005	0.006	0.008
		(2) var	0.000	0.000	0.000	0.055	1.323	0.000	0.000	0.000
	500	(1) mean	0.001	0.000	0.002	1.887	2.704	0.005	0.006	0.009
		(2) var	0.000	0.000	0.000	0.023	0.512	0.000	0.000	0.000
	1000	(1) mean	0.001	0.000	0.002	1.885	2.654	0.005	0.006	0.009
		(2) var	0.000	0.000	0.000	0.010	0.220	0.000	0.000	0.000

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
GDDM	10	(1) mean	0.004	0.004	0.000	-0.039	0.087	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.287	2.065	0.000	0.000	0.000
	50	(1) mean	0.004	0.004	0.000	0.120	0.155	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.122	0.779	0.000	0.000	0.000
	100	(1) mean	0.004	0.004	0.000	0.046	0.219	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.091	0.534	0.000	0.000	0.000
	200	(1) mean	0.004	0.004	0.000	0.120	0.350	0.004	0.004	0.005
		(2) var	0.000	0.000	0.000	0.050	0.287	0.000	0.000	0.000
	250	(1) mean	0.004	0.004	0.000	0.115	0.293	0.004	0.004	0.005
		(2) var	0.000	0.000	0.000	0.036	0.204	0.000	0.000	0.000
	500	(1) mean	0.004	0.004	0.000	0.075	0.198	0.004	0.004	0.005
		(2) var	0.000	0.000	0.000	0.016	0.113	0.000	0.000	0.000
	1000	(1) mean	0.004	0.004	0.000	0.100	0.252	0.004	0.004	0.005
		(2) var	0.000	0.000	0.000	0.009	0.065	0.000	0.000	0.000

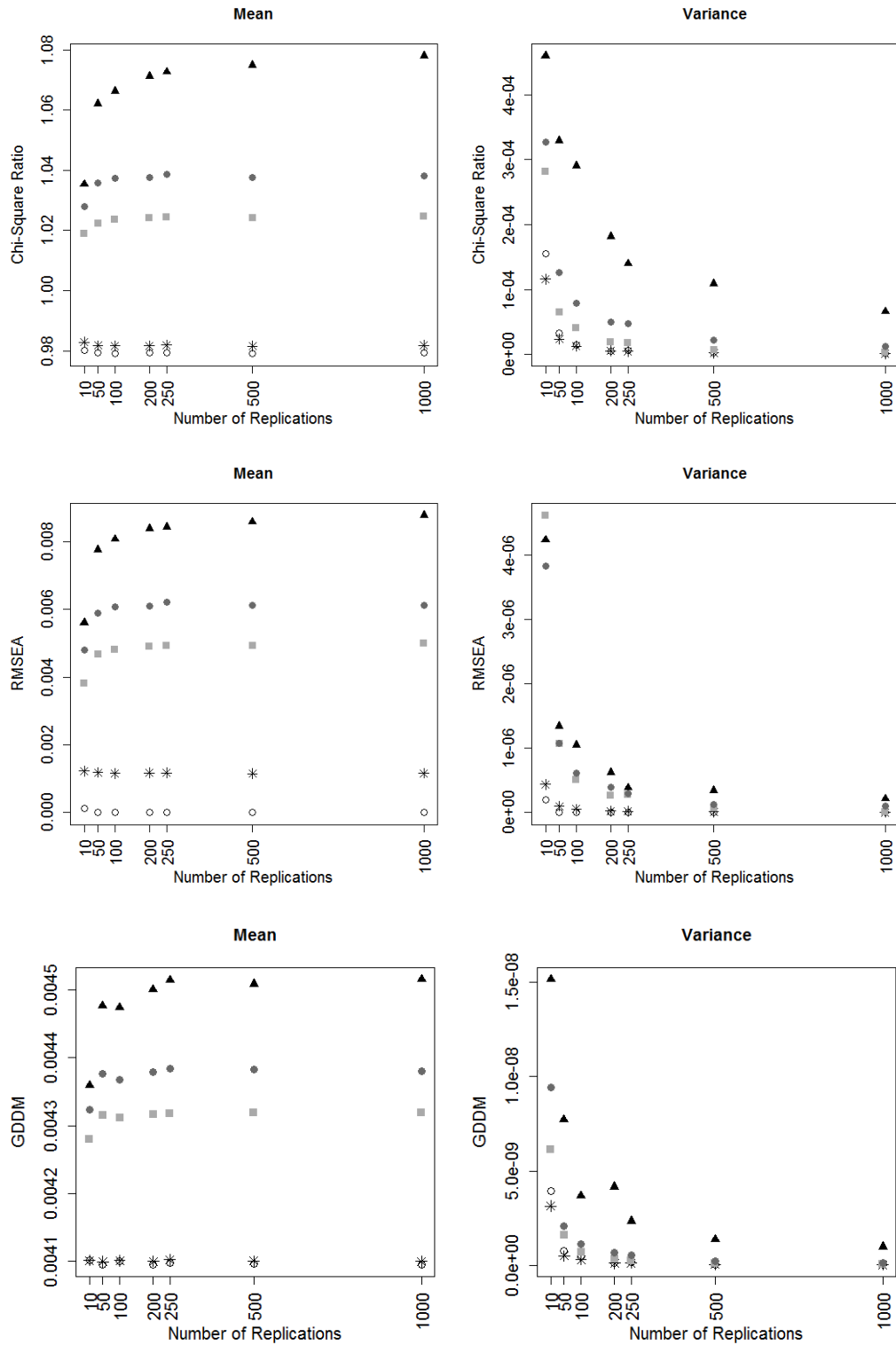


Figure C.1. Distributional and key indicators for model fit indices, 2-factor models.

Fit results for this study are presented separately for items estimated as between-item multidimensional (Table C.2 and Figure C.2) and within-item multidimensional (Table C.3 and Figure C.3). Values for the Modification Index and Wald Test are presented for only a single factor as items loading on the second factor were similar in magnitude and patterns of behavior. Point estimates of the distributional indicators achieve stability between 100 and 200 replications; for 100 replications and greater, values for the $S\text{-}\chi^2$ and Modification Index typically differ in the tenths decimal place while the Wald Test values differ at the hundredths decimal place. Point estimates of the 90th and 95th percentiles achieve stability at 100 replications while the 99th percentile is somewhat unstable across all replication sets. Representing the largest, most extreme item fit values, the precision of the key indicators is seen to increase by 100 and 250 replications; though the variance is still quite large for these indicators, proportional decreases are largest across replication sets 10, 50, and 100.

Table C.2

2-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets, Between-Item Multidimensionality

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
S- χ^2	10	(1) mean	27.7138	26.9725	7.41778	0.29765	0.13205	36.1733	38.5074	40.3748
		(2) var	5.987	7.794	4.163	0.198	1.634	21.610	25.916	34.596
	50	(1) mean	27.827	27.114	7.691	0.573	0.668	37.330	40.682	46.686
		(2) var	0.992	1.672	0.979	0.198	3.069	4.657	6.915	26.081
	100	(1) mean	27.750	26.892	7.663	0.700	1.067	37.532	40.882	47.614
		(2) var	0.642	0.984	0.623	0.193	3.787	2.422	5.535	22.907
	200	(1) mean	27.601	26.804	7.582	0.766	1.408	37.487	40.932	47.310
		(2) var	0.276	0.359	0.233	0.103	3.164	1.057	3.072	6.309
	250	(1) mean	27.731	26.865	7.747	0.772	1.309	37.759	41.748	48.175
		(2) var	0.285	0.401	0.232	0.079	1.844	0.678	2.499	7.095
	500	(1) mean	27.642	26.825	7.645	0.738	1.171	37.626	41.562	48.133
		(2) var	0.093	0.174	0.100	0.044	1.177	0.247	1.198	1.862
	1000	(1) mean	27.714	26.928	7.660	0.751	1.290	37.674	41.679	48.329
		(2) var	0.060	0.080	0.057	0.025	0.595	0.135	0.547	0.878
Mod. Index	10	(1) mean	1.984	1.321	2.270	0.301	0.568	3.921	4.351	4.694
		(2) var	2.543	1.319	4.948	0.171	7.297	12.662	16.731	20.651
	50	(1) mean	2.316	1.025	3.381	1.917	5.913	5.336	8.103	12.043
		(2) var	0.614	0.232	1.955	0.529	23.138	6.345	13.154	26.719
	100	(1) mean	2.238	0.952	3.383	2.387	7.938	5.654	8.630	13.807
		(2) var	0.253	0.116	0.884	0.482	27.486	3.785	9.986	16.417
	200	(1) mean	2.303	0.976	3.523	2.730	9.745	5.771	9.287	15.624
		(2) var	0.143	0.064	0.456	0.322	26.193	2.028	6.487	9.759
	250	(1) mean	2.268	0.908	3.510	2.729	9.398	5.865	9.438	15.999
		(2) var	0.132	0.046	0.407	0.317	20.132	2.136	6.359	12.229
	500	(1) mean	2.251	0.921	3.519	2.907	10.382	5.763	9.547	16.782
		(2) var	0.079	0.020	0.239	0.156	11.315	1.093	3.584	9.367
	1000	(1) mean	2.259	0.926	3.541	2.965	10.578	5.711	9.746	17.111
		(2) var	0.038	0.016	0.104	0.092	6.640	0.564	1.751	4.687

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
Wald Test	10	(1) mean	40.749	40.540	4.067	0.064	-0.117	44.576	45.204	45.707
		(2) var	4.575	5.334	2.834	0.177	4.514	8.326	10.149	12.133
	50	(1) mean	41.318	40.899	4.420	0.363	0.137	46.630	48.209	50.478
		(2) var	0.863	1.603	0.529	0.199	1.736	2.112	4.112	7.588
	100	(1) mean	41.145	40.605	4.459	0.462	0.054	46.825	48.520	51.523
		(2) var	0.438	0.673	0.200	0.087	0.861	1.403	1.759	4.637
	200	(1) mean	41.275	40.706	4.494	0.476	0.059	47.155	49.003	52.161
		(2) var	0.250	0.305	0.110	0.060	0.426	0.914	1.146	4.102
	250	(1) mean	41.273	40.683	4.487	0.498	0.064	47.133	49.026	52.599
		(2) var	0.163	0.216	0.079	0.040	0.337	0.649	0.837	2.742
	500	(1) mean	41.215	40.624	4.485	0.512	0.105	47.141	49.053	52.689
		(2) var	0.085	0.090	0.047	0.025	0.185	0.367	0.488	1.816
	1000	(1) mean	41.216	40.633	4.495	0.512	0.107	47.152	49.051	53.039
		(2) var	0.027	0.029	0.022	0.008	0.091	0.149	0.330	0.853

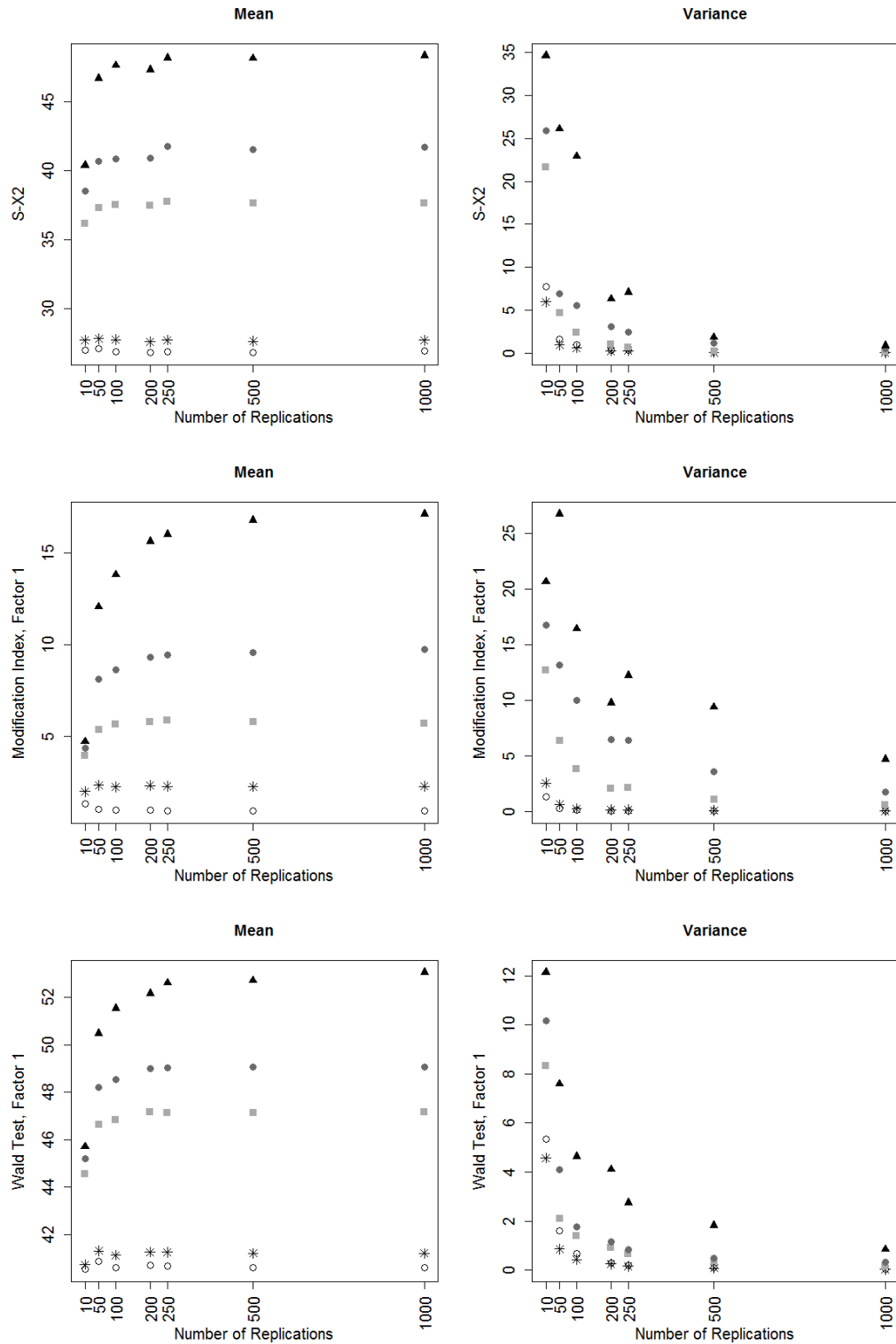


Figure C.2. Distributional and key indicators for item fit indices, 2-factor models, between-item multidimensionality.

Table C.3

2-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets, Within-Item Multidimensionality

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurtosis	90%ile	95%ile	99%ile
S- χ^2	10	(1) mean	15.222	14.661	5.473	0.439	0.754	20.996	23.358	25.247
		(2) var	2.549	3.726	2.904	0.338	4.158	11.418	16.447	27.335
	50	(1) mean	15.385	14.660	5.516	0.778	1.234	22.198	24.862	29.823
		(2) var	0.667	0.648	0.520	0.182	2.841	2.572	5.225	14.080
	100	(1) mean	15.376	14.744	5.560	0.820	1.285	22.284	25.137	31.183
		(2) var	0.307	0.318	0.275	0.089	1.010	1.489	3.404	13.034
	200	(1) mean	15.314	14.713	5.566	0.873	1.517	22.311	25.035	32.141
		(2) var	0.174	0.187	0.113	0.058	0.889	0.704	1.094	9.736
	250	(1) mean	15.302	14.649	5.597	0.900	1.480	22.315	25.284	32.671
		(2) var	0.102	0.122	0.098	0.040	0.524	0.509	0.871	6.998
	500	(1) mean	15.350	14.708	5.585	0.896	1.511	22.350	25.336	32.958
		(2) var	0.059	0.067	0.070	0.023	0.311	0.374	0.577	6.278
	1000	(1) mean	15.341	14.699	5.610	0.913	1.530	22.400	25.412	33.923
		(2) var	0.027	0.036	0.024	0.009	0.119	0.180	0.233	1.384
Wald Test	10	(1) mean	13.215	13.228	2.401	-0.008	-0.010	15.838	16.419	16.884
		(2) var	0.653	0.824	0.327	0.211	1.410	1.321	1.624	2.370
	50	(1) mean	13.214	13.272	2.417	-0.040	-0.217	16.179	16.976	18.100
		(2) var	0.135	0.182	0.060	0.078	0.243	0.307	0.370	0.807
	100	(1) mean	13.231	13.306	2.409	-0.014	-0.199	16.248	17.092	18.353
		(2) var	0.056	0.081	0.023	0.040	0.137	0.193	0.248	0.383
	200	(1) mean	13.179	13.267	2.417	-0.037	-0.296	16.230	17.054	18.420
		(2) var	0.028	0.041	0.012	0.018	0.062	0.110	0.118	0.255
	250	(1) mean	13.179	13.227	2.419	0.009	-0.259	16.277	17.101	18.567
		(2) var	0.022	0.028	0.011	0.011	0.040	0.083	0.071	0.202
	500	(1) mean	13.200	13.266	2.440	0.010	-0.282	16.351	17.193	18.693
		(2) var	0.014	0.016	0.004	0.008	0.018	0.040	0.046	0.120
	1000	(1) mean	13.198	13.269	2.425	-0.002	-0.271	16.333	17.172	18.686
		(2) var	0.005	0.010	0.003	0.003	0.008	0.025	0.014	0.040

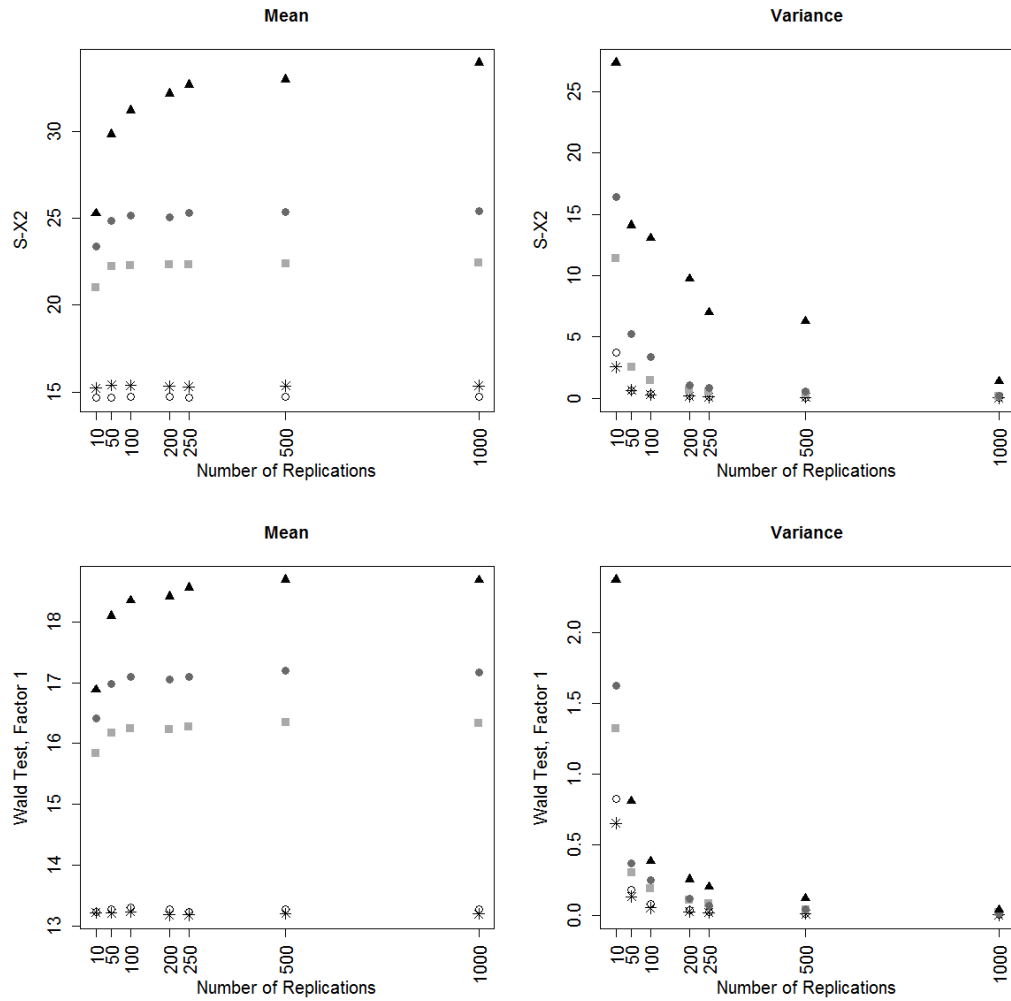


Figure C.3. Distributional and key indicators for item fit indices, 2-factor models, within-item multidimensionality.

Items estimated as within-item multidimensional present patterns comparable to those seen in the between-item multidimensional items, though it is notable that the dispersion of the values for the within-item multidimensional items is greater. Means, medians, and key indicator values are generally smaller compared to the between-item multidimensional results with similar precision, following the pattern of results seen for the between-item multidimensional items.

C.4. Replications and the Three-Factor Model

The highly-correlated 3-factor model, comprised of responses to 12 moderately-discriminating / high-difficulty (MH) between-item multidimensional items by 250 simulated examinees was identified as the second boundary condition and anticipated to yield the most unstable results. Again, 1,000 replications were divided into equally-sized partitions.

Model fit results for this model are presented in Table C.4 and Figure C.4 . Distributional indicators for the model fit indices appear to be stable and precise by 100 replications; the averages are quite stable and the variances are less than 0.001 with the largest decreases in variance occurring by 200 replications. As seen in the two-factor model, the key indicators for all three model fit indices also achieve stability by 100 replications, though the mean of the 99th percentile shows some fluctuation across replication sets, with the exception of the GDDM, again likely due to the fact that values of this index are very small.

Table C.4

3-Factor Model: Distributional and Key Indicators for Model Fit Indices Across Partition Sets

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
χ^2/df	10	(1) mean	0.969	0.955	0.146	0.263	0.095	1.130	1.177	1.214
		(2) var	0.002	0.003	0.001	0.247	2.282	0.008	0.008	0.011
	50	(1) mean	0.960	0.947	0.140	0.465	0.349	1.135	1.198	1.294
		(2) var	0.000	0.001	0.000	0.113	0.947	0.001	0.002	0.004
	100	(1) mean	0.961	0.946	0.142	0.487	0.329	1.145	1.208	1.326
		(2) var	0.000	0.000	0.000	0.076	0.414	0.001	0.002	0.005
	200	(1) mean	0.960	0.945	0.142	0.538	0.442	1.145	1.208	1.340
		(2) var	0.000	0.000	0.000	0.032	0.256	0.001	0.001	0.002
	250	(1) mean	0.960	0.944	0.143	0.544	0.399	1.149	1.211	1.341
		(2) var	0.000	0.000	0.000	0.023	0.190	0.000	0.000	0.002
	500	(1) mean	0.960	0.946	0.142	0.525	0.439	1.146	1.208	1.350
		(2) var	0.000	0.000	0.000	0.012	0.088	0.000	0.000	0.001
	1000	(1) mean	0.961	0.946	0.142	0.519	0.399	1.150	1.211	1.352
		(2) var	0.000	0.000	0.000	0.007	0.049	0.000	0.000	0.001
RMSEA	10	(1) mean	0.008	0.002	0.011	0.952	1.047	0.021	0.025	0.028
		(2) var	0.000	0.000	0.000	0.339	11.280	0.000	0.000	0.000
	50	(1) mean	0.007	0.000	0.010	1.263	0.727	0.023	0.028	0.034
		(2) var	0.000	0.000	0.000	0.098	1.425	0.000	0.000	0.000
	100	(1) mean	0.007	0.000	0.011	1.303	0.631	0.024	0.029	0.036
		(2) var	0.000	0.000	0.000	0.050	0.570	0.000	0.000	0.000
	200	(1) mean	0.007	0.000	0.011	1.351	0.744	0.024	0.029	0.037
		(2) var	0.000	0.000	0.000	0.035	0.398	0.000	0.000	0.000
	250	(1) mean	0.007	0.000	0.011	1.345	0.677	0.024	0.029	0.037
		(2) var	0.000	0.000	0.000	0.024	0.264	0.000	0.000	0.000
	500	(1) mean	0.007	0.000	0.011	1.347	0.666	0.024	0.029	0.037
		(2) var	0.000	0.000	0.000	0.014	0.160	0.000	0.000	0.000

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
GDDM	1000	(1) mean	0.007	0.000	0.011	1.331	0.581	0.025	0.029	0.037
		(2) var	0.000	0.000	0.000	0.006	0.069	0.000	0.000	0.000
	10	(1) mean	0.003	0.003	0.000	-0.076	0.072	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.218	1.623	0.000	0.000	0.000
	50	(1) mean	0.003	0.003	0.000	-0.137	-0.001	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.101	0.622	0.000	0.000	0.000
	100	(1) mean	0.003	0.003	0.000	-0.129	-0.052	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.056	0.381	0.000	0.000	0.000
	200	(1) mean	0.003	0.003	0.000	-0.090	-0.012	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.029	0.105	0.000	0.000	0.000
	250	(1) mean	0.003	0.003	0.000	-0.082	-0.041	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.017	0.096	0.000	0.000	0.000
	500	(1) mean	0.003	0.003	0.000	-0.106	-0.005	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.012	0.054	0.000	0.000	0.000
	1000	(1) mean	0.003	0.003	0.000	-0.116	-0.047	0.004	0.004	0.004
		(2) var	0.000	0.000	0.000	0.006	0.019	0.000	0.000	0.000

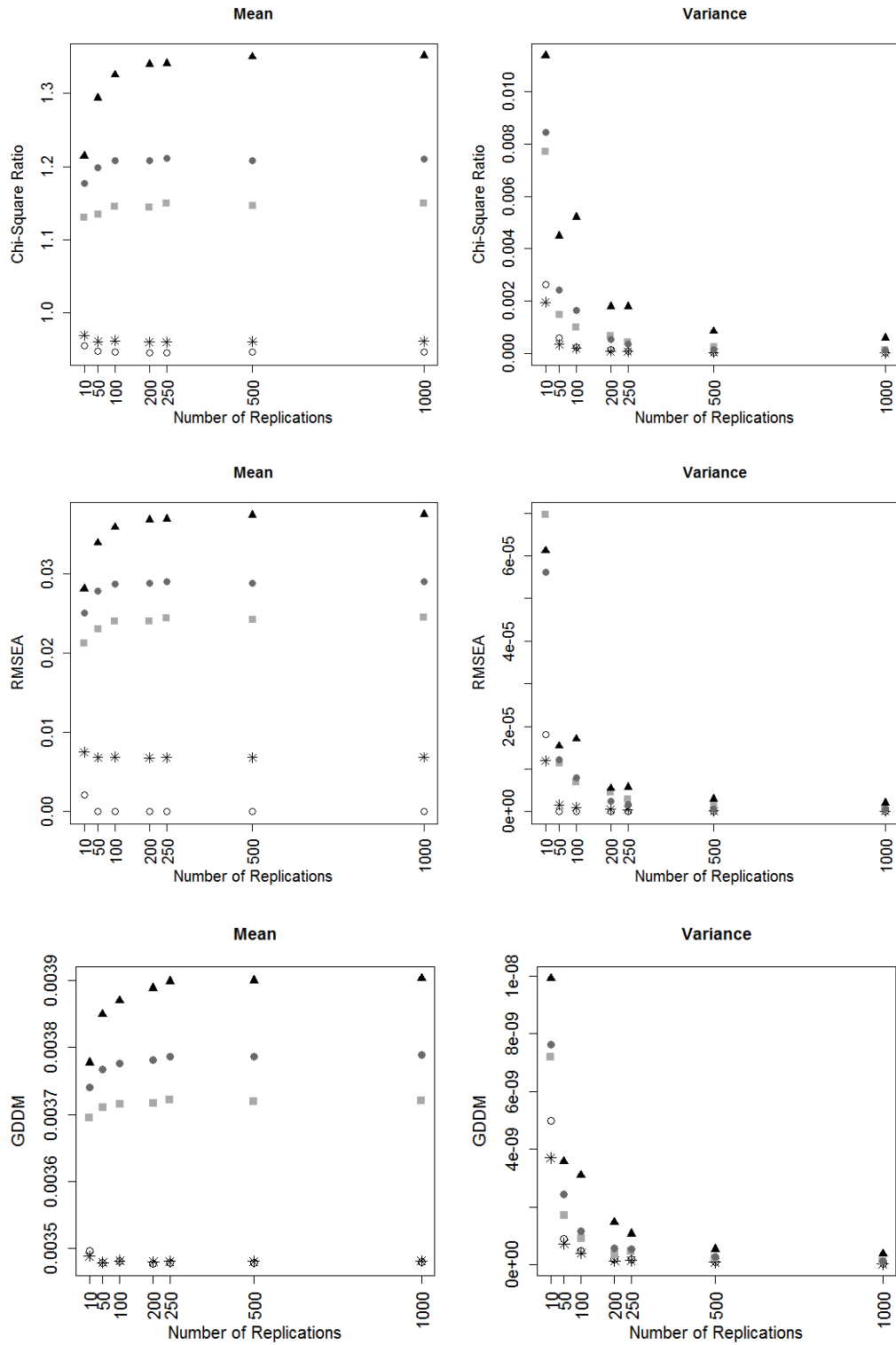


Figure C.4. Distributional and key indicators for model fit indices, 3-factor models.

Item fit results for the 3-factor model are presented in Table C.5 and Figure C.5. Distributional indicators for the $S-\chi^2$ are stable at 100 replications and the greatest gains in precision are also achieved at 100 replications. The key indicators are less stable and precise, as expected, though the means and variances show the greatest improvements between 100 and 200 replications. Though the results for the Modification Index and Wald test show similar patterns for the distributional indicators, the key indicators for these statistics under a 3-factor model show greater instability and imprecision across replication sets. While the means and variances of the 90th and 95th percentiles for the item fit indices are relatively stable and precise by 100 replications, means and variances of the 99th percentiles show fluctuation across replication sets and large decreases in variances for replication sets of 250 replications and larger.

Table C.5

3-Factor Model: Distributional and Key Indicators for Item Fit Indices Across Partition Sets

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
S- χ^2	10	(1) mean	7.099	6.659	3.127	0.504	0.885	10.411	11.821	12.949
		(2) var	0.977	1.315	1.066	0.353	4.312	3.459	6.117	10.830
	50	(1) mean	7.227	6.833	3.336	0.850	1.453	11.357	13.029	16.097
		(2) var	0.201	0.339	0.192	0.216	4.580	1.078	1.884	5.093
	100	(1) mean	7.157	6.756	3.293	0.959	1.899	11.226	12.959	16.584
		(2) var	0.091	0.133	0.103	0.155	2.837	0.627	1.273	5.428
	200	(1) mean	7.125	6.740	3.329	1.031	2.042	11.262	13.165	17.443
		(2) var	0.063	0.091	0.059	0.076	1.512	0.392	0.706	4.748
	250	(1) mean	7.114	6.749	3.287	0.994	1.839	11.252	13.018	17.197
		(2) var	0.053	0.050	0.052	0.059	1.059	0.332	0.654	3.646
	500	(1) mean	7.146	6.751	3.300	1.059	2.089	11.269	13.223	17.623
		(2) var	0.020	0.031	0.027	0.035	0.577	0.126	0.371	2.715
	1000	(1) mean	7.135	6.748	3.312	1.080	2.188	11.281	13.175	17.824
		(2) var	0.009	0.014	0.013	0.018	0.303	0.081	0.116	1.855
Mod. Index	10	(1) mean	0.800	0.524	0.907	0.643	1.094	1.779	2.075	2.311
		(2) var	0.193	0.279	0.259	0.245	6.919	0.831	1.198	1.612
	50	(1) mean	0.839	0.388	1.206	2.073	6.529	2.142	2.909	4.692
		(2) var	0.054	0.026	0.243	0.832	45.069	0.450	0.804	4.963
	100	(1) mean	0.850	0.385	1.211	2.377	8.186	2.188	3.073	5.185
		(2) var	0.025	0.010	0.109	0.750	54.749	0.204	0.570	2.811
	200	(1) mean	0.861	0.386	1.289	3.149	16.122	2.238	3.079	5.488
		(2) var	0.014	0.006	0.083	1.600	197.872	0.107	0.311	2.099
	250	(1) mean	0.861	0.375	1.331	3.326	17.389	2.251	3.115	6.098
		(2) var	0.010	0.003	0.064	1.471	206.422	0.073	0.259	2.209
	500	(1) mean	0.847	0.379	1.279	3.441	19.374	2.240	3.097	5.966
		(2) var	0.006	0.002	0.032	1.139	192.483	0.049	0.168	1.273

Fit Index	Split	Statistic	Mean	Median	SD	Skew	Kurt.	90%ile	95%ile	99%ile
Wald Test	1000	(1) mean	0.845	0.380	1.276	3.705	23.317	2.240	3.037	6.023
		(2) var	0.002	0.001	0.018	0.865	150.121	0.016	0.064	1.125
	10	(1) mean	8.210	8.033	2.056	0.102	0.448	9.768	10.035	10.248
		(2) var	2.597	2.733	1.218	0.114	4.496	4.888	5.619	6.301
	50	(1) mean	8.314	8.154	2.471	0.263	0.525	11.162	12.070	13.063
		(2) var	0.489	0.634	0.218	0.279	1.993	1.429	1.997	2.950
	100	(1) mean	8.273	8.162	2.440	0.317	0.325	11.295	12.166	13.573
		(2) var	0.180	0.280	0.124	0.167	1.175	0.750	1.009	1.844
	200	(1) mean	8.184	8.050	2.405	0.354	0.418	11.199	12.169	14.014
		(2) var	0.085	0.131	0.060	0.129	0.827	0.258	0.702	1.497
	250	(1) mean	8.242	8.099	2.401	0.382	0.432	11.258	12.246	14.173
		(2) var	0.076	0.075	0.032	0.089	0.634	0.189	0.426	1.041
	500	(1) mean	8.215	8.056	2.382	0.393	0.393	11.257	12.291	14.323
		(2) var	0.031	0.038	0.020	0.043	0.286	0.130	0.253	0.678
	1000	(1) mean	8.250	8.087	2.417	0.412	0.341	11.296	12.458	14.604
		(2) var	0.015	0.012	0.009	0.022	0.126	0.074	0.149	0.254

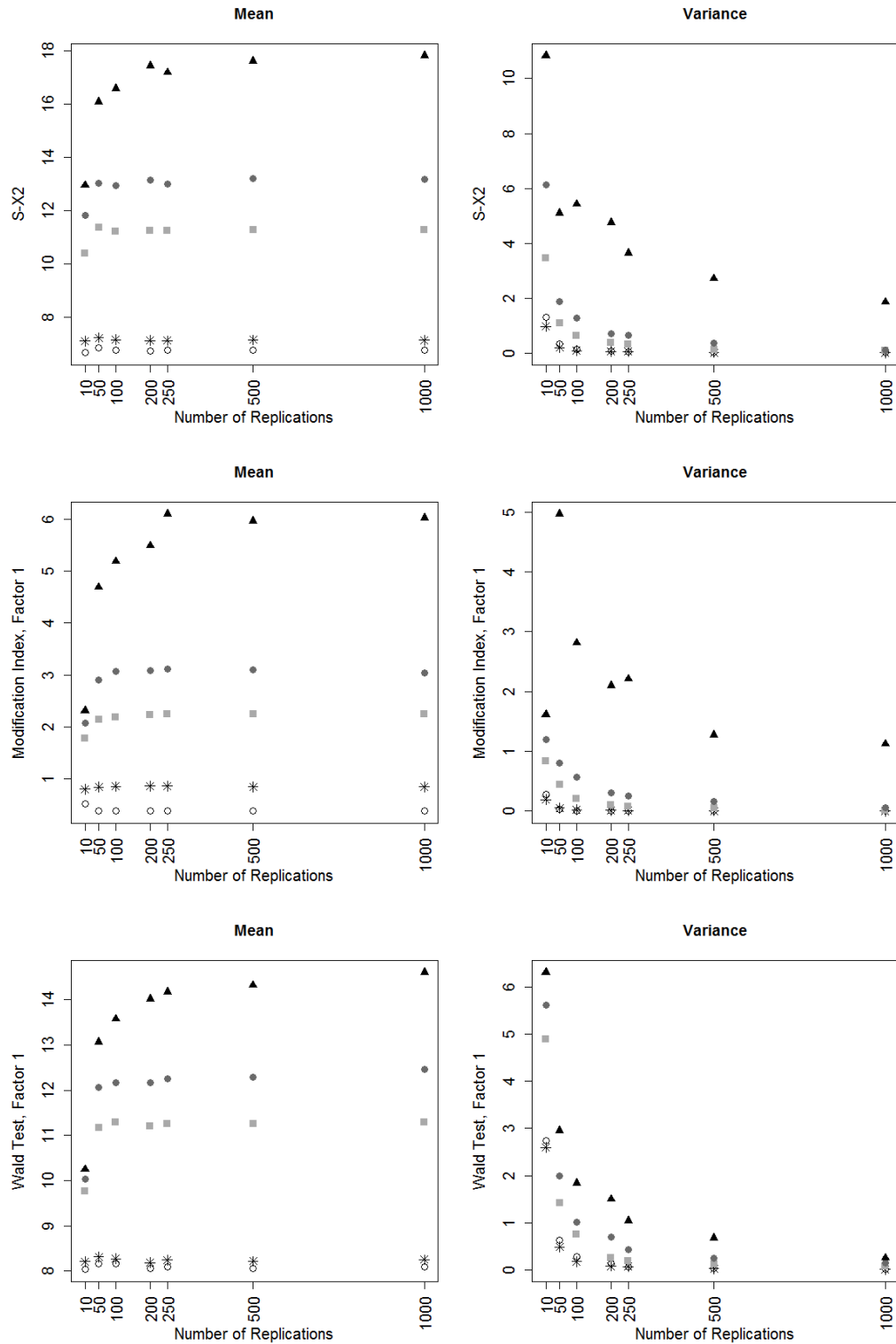


Figure C.5. Distributional and key indicators for item fit indices, 3-factor models.

Overall, fluctuations and instability in the $S-\chi^2$ is typically the result of large item-specific estimates, especially under the 2-factor model where the mean was approximately 27. The magnitude of differences between means and variances for each replication set are, therefore, greater than those seen for other indices. Similarly, the stability and precision of the Wald Test must be considered in the context of large item-specific values. Instability of the Modification Index is owed to the positive skewness (typically greater than 2.0) under each replication set, indicating a distribution with a long tail containing a few exceptionally large positive values.

The results of this study suggest that model- and item-fit indices achieve an acceptable level of precision and stability at 100 to 200 replications, though some exceptions exist. Means, medians, and standard deviations for all indices were demonstrated to be extremely precise and stable across all levels of replication; estimates of 90th, 95th, and 99th percentiles evidence less precision and stability owing to the extreme nature of these values.

Appendix D

Key Descriptive Statistics Under Misspecified Model Estimation

Table D.1

Key Descriptive Statistics for the χ^2/df Model-Fit Index Under Misspecified Model Estimation

Item Type	Sample Size	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
HH	250	H	0.797	1.105	1.265	1.185	1.338	1.572	1.771	2.253	3.053	0.252	2.253	6.903
HH	250	M	0.966	1.406	1.787	1.613	1.991	2.509	2.934	3.736	5.441	0.548	1.842	4.314
HH	250	L	1.110	1.776	2.429	2.166	2.785	3.652	4.359	5.575	7.785	0.907	1.658	3.173
HH	1000	H	1.026	1.587	1.986	1.773	2.280	2.826	3.192	3.916	5.296	0.579	1.468	2.164
HH	1000	M	1.980	3.188	4.404	3.778	5.294	6.738	8.035	9.862	13.024	1.685	1.414	1.800
HH	1000	L	3.205	5.339	7.593	6.486	9.043	11.716	15.042	17.748	20.820	3.168	1.488	2.004
HM	250	H	0.833	1.085	1.232	1.160	1.300	1.516	1.685	2.093	4.914	0.232	3.016	21.125
HM	250	M	0.866	1.404	1.830	1.650	2.062	2.660	3.069	3.923	5.345	0.597	1.655	3.204
HM	250	L	1.320	1.930	2.741	2.392	3.251	4.390	5.118	6.367	8.902	1.133	1.389	1.814
HM	1000	H	1.123	1.558	2.100	1.890	2.377	3.196	3.645	4.463	6.547	0.735	1.417	1.904
HM	1000	M	1.979	3.424	5.281	4.584	6.267	9.114	10.420	12.544	14.662	2.442	1.201	0.895
HM	1000	L	3.600	6.291	9.794	8.565	11.736	16.609	20.292	23.547	28.403	4.722	1.176	0.846
HL	250	H	0.784	1.130	1.340	1.243	1.443	1.727	1.987	2.497	3.766	0.309	2.060	5.701
HL	250	M	1.165	1.622	2.250	2.016	2.583	3.394	4.098	5.138	7.353	0.847	1.541	2.572
HL	250	L	1.547	2.366	3.560	3.161	4.246	5.837	6.970	8.532	11.328	1.574	1.338	1.588
HL	1000	H	1.370	2.044	2.886	2.510	3.489	4.481	5.219	6.412	8.800	1.127	1.302	1.530
HL	1000	M	2.942	4.962	7.592	6.634	9.440	12.373	15.033	17.647	21.367	3.430	1.114	0.752
HL	1000	L	4.595	8.606	13.486	11.984	16.591	22.491	28.081	31.673	37.884	6.376	1.130	0.713
MH	250	H	0.640	1.041	1.110	1.093	1.162	1.263	1.360	1.589	2.444	0.140	1.346	6.241
MH	250	M	0.710	1.249	1.446	1.368	1.561	1.828	2.028	2.543	3.574	0.305	1.777	5.110
MH	250	L	0.932	1.530	1.902	1.747	2.119	2.607	2.973	3.856	5.433	0.543	1.782	4.362
MH	1000	H	0.648	1.415	1.672	1.574	1.854	2.177	2.385	2.944	4.028	0.382	1.403	2.941

Item Type	Sample Size	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
MH	1000	M	1.410	2.560	3.380	3.048	3.947	4.938	5.672	7.067	9.066	1.121	1.363	1.993
MH	1000	L	2.539	4.025	5.597	4.964	6.693	8.559	10.054	12.444	16.616	2.122	1.318	1.663
MM	250	H	0.639	1.046	1.134	1.103	1.198	1.314	1.422	1.693	2.489	0.158	1.573	6.070
MM	250	M	0.811	1.302	1.567	1.465	1.733	2.070	2.294	2.868	4.457	0.376	1.595	3.688
MM	250	L	1.084	1.664	2.173	1.974	2.481	3.134	3.590	4.627	5.872	0.706	1.515	2.705
MM	1000	H	0.783	1.499	1.835	1.720	2.080	2.463	2.738	3.243	4.279	0.460	1.224	1.934
MM	1000	M	1.884	2.887	3.943	3.564	4.695	5.962	6.755	8.156	9.920	1.391	1.098	0.862
MM	1000	L	2.650	4.750	6.794	6.051	8.251	10.747	12.137	14.455	18.097	2.664	1.030	0.635
ML	250	H	0.544	1.089	1.202	1.161	1.272	1.434	1.585	1.855	2.390	0.188	1.450	3.835
ML	250	M	0.942	1.412	1.747	1.620	1.948	2.373	2.672	3.374	5.200	0.467	1.590	3.486
ML	250	L	1.290	1.863	2.507	2.291	2.901	3.672	4.301	5.429	6.732	0.863	1.436	2.314
ML	1000	H	0.906	1.651	2.073	1.930	2.374	2.856	3.163	3.897	5.134	0.563	1.213	1.695
ML	1000	M	2.222	3.316	4.691	4.267	5.658	7.113	8.194	9.919	12.408	1.735	1.080	0.888
ML	1000	L	3.274	5.470	8.169	7.383	10.062	12.801	15.158	18.075	20.618	3.345	1.019	0.644

Table D.2

Key Descriptive Statistics for the RMSEA Model-Fit Index Under Misspecified Model Estimation

Item Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
HH	2	H	0.000	0.027	0.035	0.034	0.042	0.049	0.056	0.070	0.091	0.012	0.595	1.143
HH	2	M	0.016	0.051	0.064	0.062	0.073	0.086	0.092	0.105	0.133	0.016	0.517	0.143
HH	2	L	0.043	0.072	0.087	0.086	0.099	0.117	0.125	0.137	0.165	0.020	0.557	-0.113
HH	3	H	0.000	0.021	0.025	0.024	0.028	0.034	0.039	0.051	0.089	0.008	0.912	4.833
HH	3	M	0.000	0.041	0.047	0.046	0.052	0.059	0.064	0.073	0.126	0.009	0.687	2.072
HH	3	L	0.021	0.056	0.064	0.063	0.072	0.081	0.086	0.096	0.124	0.012	0.458	0.382
HM	2	H	0.000	0.027	0.036	0.035	0.044	0.051	0.056	0.066	0.115	0.012	0.509	0.506
HM	2	M	0.027	0.055	0.070	0.068	0.083	0.095	0.102	0.111	0.132	0.018	0.251	-0.587
HM	2	L	0.049	0.082	0.100	0.099	0.118	0.134	0.142	0.153	0.178	0.024	0.185	-0.649
HM	3	H	0.000	0.018	0.023	0.023	0.028	0.033	0.037	0.047	0.125	0.008	0.751	5.513
HM	3	M	0.000	0.040	0.048	0.048	0.056	0.062	0.067	0.073	0.091	0.011	0.164	-0.378
HM	3	L	0.036	0.059	0.070	0.070	0.080	0.088	0.093	0.102	0.117	0.014	0.088	-0.551
HL	2	H	0.000	0.036	0.046	0.045	0.054	0.064	0.069	0.079	0.105	0.013	0.401	-0.079
HL	2	M	0.040	0.070	0.086	0.084	0.100	0.115	0.123	0.133	0.159	0.020	0.304	-0.544
HL	2	L	0.062	0.100	0.121	0.117	0.138	0.160	0.168	0.178	0.203	0.026	0.349	-0.637
HL	3	H	0.000	0.023	0.030	0.030	0.036	0.042	0.046	0.055	0.083	0.010	0.256	1.070
HL	3	M	0.026	0.050	0.060	0.059	0.070	0.079	0.084	0.093	0.109	0.014	0.352	-0.455
HL	3	L	0.047	0.072	0.085	0.083	0.098	0.109	0.116	0.125	0.144	0.017	0.366	-0.571
MH	2	H	0.000	0.021	0.026	0.026	0.032	0.037	0.041	0.049	0.076	0.010	-0.297	1.097
MH	2	M	0.000	0.042	0.051	0.050	0.058	0.066	0.072	0.082	0.101	0.012	0.241	0.631
MH	2	L	0.026	0.060	0.071	0.070	0.081	0.093	0.100	0.112	0.133	0.016	0.447	0.064
MH	3	H	0.000	0.015	0.018	0.020	0.023	0.027	0.030	0.039	0.062	0.008	-0.461	0.834
MH	3	M	0.000	0.032	0.037	0.037	0.042	0.048	0.052	0.059	0.083	0.009	-0.619	2.822
MH	3	L	0.000	0.046	0.053	0.052	0.059	0.065	0.071	0.078	0.099	0.010	0.240	0.851
MM	2	H	0.000	0.023	0.029	0.029	0.035	0.041	0.045	0.053	0.077	0.010	-0.311	1.105
MM	2	M	0.013	0.048	0.058	0.057	0.066	0.075	0.080	0.088	0.118	0.013	0.244	-0.016

Item Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
MM	2	L	0.041	0.070	0.082	0.081	0.093	0.104	0.111	0.123	0.140	0.017	0.280	-0.172
MM	3	H	0.000	0.016	0.020	0.021	0.025	0.029	0.032	0.040	0.059	0.009	-0.391	0.643
MM	3	M	0.000	0.035	0.041	0.041	0.046	0.052	0.056	0.064	0.089	0.009	-0.216	1.855
MM	3	L	0.018	0.051	0.058	0.058	0.065	0.072	0.076	0.085	0.102	0.011	0.238	0.211
ML	2	H	0.000	0.028	0.034	0.034	0.040	0.047	0.051	0.059	0.075	0.010	-0.202	1.361
ML	2	M	0.007	0.056	0.065	0.064	0.074	0.083	0.089	0.100	0.130	0.014	0.349	0.148
ML	2	L	0.043	0.079	0.092	0.091	0.103	0.117	0.125	0.135	0.151	0.017	0.432	-0.146
ML	3	H	0.000	0.020	0.024	0.024	0.028	0.033	0.037	0.047	0.066	0.009	-0.320	1.665
ML	3	M	0.000	0.040	0.046	0.046	0.052	0.059	0.064	0.072	0.093	0.010	0.223	1.243
ML	3	L	0.034	0.057	0.065	0.064	0.072	0.081	0.087	0.096	0.121	0.012	0.524	0.236

Table D.3

Key Descriptive Statistics for the GDDM Model-Fit Index Under Misspecified Model Estimation

Item Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
HH	2	H	0.003	0.005	0.006	0.005	0.006	0.007	0.007	0.009	0.015	0.001	1.726	7.479
HH	2	M	0.005	0.006	0.008	0.007	0.009	0.010	0.011	0.013	0.020	0.002	1.220	3.489
HH	2	L	0.006	0.008	0.010	0.009	0.011	0.013	0.014	0.017	0.026	0.002	1.071	1.343
HH	3	H	0.003	0.004	0.004	0.004	0.005	0.005	0.006	0.006	0.007	0.001	0.304	-0.141
HH	3	M	0.003	0.005	0.006	0.005	0.006	0.007	0.007	0.008	0.009	0.001	0.606	0.349
HH	3	L	0.004	0.006	0.007	0.007	0.007	0.008	0.009	0.009	0.010	0.001	0.555	0.045
HM	2	H	0.004	0.006	0.007	0.007	0.008	0.009	0.010	0.014	0.019	0.002	2.041	6.706
HM	2	M	0.006	0.009	0.010	0.010	0.011	0.012	0.013	0.017	0.023	0.002	1.427	4.607
HM	2	L	0.008	0.011	0.013	0.013	0.015	0.017	0.019	0.022	0.030	0.003	0.975	1.310
HM	3	H	0.003	0.004	0.005	0.005	0.006	0.006	0.006	0.007	0.008	0.001	-0.183	-0.492
HM	3	M	0.004	0.006	0.006	0.006	0.007	0.007	0.008	0.008	0.010	0.001	0.058	-0.136
HM	3	L	0.004	0.007	0.008	0.008	0.008	0.009	0.010	0.010	0.012	0.001	0.315	0.070
HL	2	H	0.005	0.007	0.008	0.008	0.009	0.010	0.011	0.014	0.019	0.002	1.381	4.158
HL	2	M	0.008	0.011	0.013	0.013	0.014	0.016	0.017	0.019	0.025	0.002	0.688	0.735
HL	2	L	0.011	0.014	0.017	0.016	0.020	0.023	0.024	0.026	0.032	0.004	0.674	-0.445
HL	3	H	0.004	0.005	0.006	0.006	0.007	0.008	0.008	0.009	0.011	0.001	0.132	-0.812
HL	3	M	0.005	0.007	0.008	0.008	0.009	0.010	0.010	0.011	0.013	0.001	0.179	-0.627
HL	3	L	0.006	0.009	0.010	0.010	0.011	0.012	0.013	0.014	0.015	0.002	0.062	-0.959
MH	2	H	0.004	0.005	0.007	0.007	0.008	0.009	0.009	0.012	0.016	0.002	1.021	3.162
MH	2	M	0.005	0.007	0.008	0.008	0.009	0.010	0.011	0.012	0.020	0.002	0.821	2.842
MH	2	L	0.006	0.008	0.010	0.009	0.011	0.013	0.013	0.015	0.022	0.002	0.644	0.218
MH	3	H	0.004	0.005	0.006	0.006	0.007	0.007	0.008	0.009	0.012	0.001	0.038	-0.312
MH	3	M	0.005	0.006	0.007	0.007	0.008	0.008	0.009	0.009	0.011	0.001	0.263	-0.384
MH	3	L	0.004	0.007	0.008	0.008	0.009	0.009	0.010	0.011	0.012	0.001	0.566	0.063
MM	2	H	0.005	0.006	0.008	0.008	0.009	0.010	0.011	0.013	0.223	0.004	40.679	2012.237
MM	2	M	0.006	0.008	0.010	0.010	0.011	0.012	0.013	0.015	0.022	0.002	0.601	1.575

Item Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
MM	2	L	0.007	0.010	0.012	0.012	0.014	0.015	0.016	0.018	0.026	0.002	0.525	0.140
MM	3	H	0.004	0.006	0.007	0.007	0.008	0.008	0.008	0.009	0.012	0.001	-0.273	-0.763
MM	3	M	0.005	0.007	0.008	0.008	0.008	0.009	0.009	0.010	0.012	0.001	-0.112	-0.395
MM	3	L	0.006	0.008	0.009	0.009	0.010	0.010	0.011	0.012	0.014	0.001	0.100	-0.083
ML	2	H	0.005	0.007	0.009	0.009	0.010	0.012	0.013	0.015	0.249	0.004	44.307	2837.669
ML	2	M	0.007	0.010	0.012	0.012	0.013	0.014	0.015	0.017	0.027	0.002	0.394	0.313
ML	2	L	0.009	0.012	0.015	0.014	0.017	0.019	0.020	0.022	0.027	0.003	0.455	-0.519
ML	3	H	0.005	0.006	0.008	0.008	0.009	0.010	0.010	0.011	0.013	0.002	-0.067	-1.013
ML	3	M	0.006	0.008	0.009	0.009	0.010	0.011	0.012	0.013	0.014	0.001	0.067	-0.597
ML	3	L	0.007	0.010	0.011	0.011	0.012	0.013	0.013	0.014	0.016	0.002	0.027	-0.447

Table D.4

Key Descriptive Statistics for the $S\text{-}\chi^2/\text{df}$ Item-Fit Index Under Misspecified Model Estimation

Miss. Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
mod.1.same	2	H	0.748	14.014	23.465	21.500	30.648	39.147	45.100	64.045	485.515	13.171	2.419	27.817
mod.1.same	2	M	0.288	12.043	19.338	18.137	25.191	31.917	36.235	46.744	394.960	10.067	2.269	34.697
mod.1.same	2	L	0.019	10.580	17.132	16.048	22.366	28.547	32.573	41.612	377.607	9.216	2.799	50.131
mod.1.same	3	H	0.257	13.827	22.943	20.693	29.867	39.170	45.611	61.612	112.091	12.308	1.106	1.876
mod.1.same	3	M	0.128	12.372	20.023	18.560	26.061	33.698	38.925	50.517	109.151	10.314	0.886	1.180
mod.1.same	3	L	0.020	11.279	18.531	17.216	24.321	31.609	36.376	46.423	87.870	9.706	0.804	0.827
mod.1.switch	2	H	1.051	18.973	33.159	27.897	40.930	60.721	76.713	107.132	175.100	20.947	1.662	3.574
mod.1.switch	2	M	0.487	18.904	36.535	28.279	43.396	71.487	98.827	143.678	262.609	28.094	2.156	5.792
mod.1.switch	2	L	0.345	18.680	38.908	28.599	45.400	78.869	114.221	168.741	302.292	33.249	2.333	6.573
mod.1.switch	3	H	0.894	14.434	25.193	21.871	33.861	45.398	51.610	63.412	104.602	13.880	0.883	0.443
mod.1.switch	3	M	0.551	12.384	20.241	18.627	26.857	34.235	38.641	48.059	102.568	10.176	0.735	0.489
mod.1.switch	3	L	0.066	10.909	17.448	16.338	22.883	29.252	33.149	41.379	78.703	8.679	0.697	0.500
mod.2.same	2	H	0.219	13.770	125.725	22.163	46.997	258.663	609.841	1812.885	27021.276	441.816	12.792	352.564
mod.2.same	2	M	0.116	10.579	52.242	16.111	24.678	60.954	186.305	836.792	8545.377	210.272	15.099	353.989
mod.2.same	2	L	0.169	9.540	36.172	14.583	21.375	35.270	93.543	512.590	9566.105	160.495	22.045	778.277
mod.2.same	3	H	0.308	11.145	23.281	16.977	24.424	34.249	50.723	166.019	1115.139	38.028	10.923	170.145
mod.2.same	3	M	0.063	9.599	19.738	14.858	21.287	28.961	37.868	142.120	850.963	32.115	10.499	146.042
mod.2.same	3	L	0.216	8.648	18.346	13.781	20.465	28.412	35.104	117.972	759.482	29.303	10.564	145.485
sev.1.same	2	H	0.487	12.808	20.929	19.375	27.331	34.767	39.547	51.848	256.565	11.134	2.039	18.214
sev.1.same	2	M	0.292	12.029	19.323	18.094	25.177	31.987	36.422	46.876	223.324	9.896	1.472	10.703
sev.1.same	2	L	0.057	11.442	18.608	17.290	24.167	31.324	36.202	46.950	160.560	9.721	1.129	4.074
sev.1.same	3	H	0.341	13.815	22.249	20.298	28.731	37.337	42.995	58.703	126.347	11.589	1.156	2.594
sev.1.same	3	M	0.256	12.605	19.808	18.578	25.786	32.842	37.349	46.717	89.693	9.665	0.728	0.678
sev.1.same	3	L	0.237	11.653	18.828	17.502	24.540	31.849	36.662	47.022	104.341	9.728	0.848	1.113
sev.1.switch	2	H	0.555	14.594	24.221	21.340	30.858	42.378	51.020	69.023	118.980	13.581	1.245	2.009
sev.1.switch	2	M	0.716	14.647	25.236	21.699	31.906	45.547	55.720	76.167	181.082	15.252	1.514	3.469

Miss. Type	Dim.	Corr.	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
sev.1.switch	2	L	0.383	14.428	26.234	21.816	32.771	48.455	61.153	90.836	210.926	17.611	1.877	5.257
sev.1.switch	3	H	0.023	12.684	22.273	19.682	29.570	39.727	46.237	58.746	107.643	12.491	1.003	1.077
sev.1.switch	3	M	0.141	10.959	18.216	16.797	24.138	31.199	35.461	44.120	94.217	9.429	0.773	0.629
sev.1.switch	3	L	0.006	9.581	15.808	14.743	20.909	26.896	30.639	38.361	79.900	8.174	0.741	0.683
sev.1.under	2	H	1.074	17.267	32.311	26.576	41.167	59.746	70.317	110.733	266.057	22.507	2.635	13.728
sev.1.under	2	M	1.915	15.634	27.523	23.643	35.644	49.677	56.892	77.783	224.509	16.503	1.912	9.006
sev.1.under	2	L	1.581	14.518	24.692	21.616	32.033	43.939	50.602	63.724	213.732	13.950	1.594	7.146
sev.1.under	3	H	1.573	15.254	30.655	23.724	43.684	57.771	65.842	91.402	122.018	19.425	1.066	0.832
sev.1.under	3	M	1.014	12.940	24.641	20.478	34.122	44.001	50.278	77.761	122.946	15.096	1.337	2.763
sev.1.under	3	L	0.673	11.167	20.497	17.783	27.208	35.454	41.743	69.777	136.841	12.739	1.752	5.598
sev.2.under	2	H	0.104	14.654	258.076	28.113	137.582	630.501	1245.783	3545.207	25650.776	833.805	9.667	148.726
sev.2.under	2	M	0.244	11.258	89.472	18.693	38.651	174.465	412.608	1302.873	11114.762	313.766	12.359	249.235
sev.2.under	2	L	0.187	9.739	45.737	15.609	25.612	73.307	173.169	606.862	8238.769	155.991	16.747	528.486
sev.2.under	3	H	0.383	11.592	60.316	17.880	29.049	93.866	245.155	1013.244	2966.346	184.734	7.362	67.712
sev.2.under	3	M	0.707	11.695	71.013	17.420	27.518	117.722	317.521	1257.509	2901.645	214.683	6.008	42.606
sev.2.under	3	L	0.533	12.290	74.911	18.570	31.213	160.476	372.017	1084.404	2537.487	193.634	4.907	28.110

Table D.5

Key Descriptive Statistics for Modification Index 1 Under Misspecified Model Estimation

Corr.	Sample		Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
	Size	Dim.												
H	250	2	0	0.202	1.960	0.811	2.242	4.843	7.680	17.191	999.000	4.758	108.522	22441.240
H	250	3	0	0.106	1.086	0.465	1.369	2.900	4.263	7.869	38.018	1.655	3.518	21.060
H	1000	2	0	1.322	8.731	3.676	8.527	20.597	38.001	82.686	220.633	15.650	4.141	22.161
H	1000	3	0	0.262	3.205	1.124	3.457	8.337	14.187	29.278	84.052	5.711	3.795	19.278
M	250	2	0	0.689	5.470	2.292	5.619	12.943	22.819	50.937	999.000	11.927	31.789	2479.710
M	250	3	0	0.204	2.326	0.909	2.753	6.180	9.561	19.029	73.667	3.824	3.719	20.974
M	1000	2	0	5.708	27.726	11.482	23.778	66.156	129.919	257.235	999.000	48.995	3.846	18.245
M	1000	3	0	0.481	9.211	2.372	8.645	24.811	47.543	91.987	999.000	18.331	4.258	48.841
L	250	2	0	1.566	10.450	4.384	10.100	24.873	45.890	98.091	999.000	19.795	9.602	332.279
L	250	3	0	0.287	3.877	1.325	4.193	10.284	17.050	35.383	122.883	6.952	3.853	20.546
L	1000	2	0	12.033	52.968	21.580	42.221	127.962	258.624	485.490	999.000	92.389	3.599	14.881
L	1000	3	0	0.611	16.778	3.330	15.458	46.895	90.985	171.267	999.000	34.257	3.721	20.257

Table D.6
Key Descriptive Statistics for Wald Test 1 Under Misspecified Model Estimation

Item Type	Sample Size	Miss. Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
HH	250	mod.1.same	0.307	18.577	14.939	14.576	11.060	8.183	6.045	2.300	56.130	5.760	0.335	0.514
HH	250	mod.1.switch	-0.023	14.485	11.918	11.413	8.791	6.734	5.691	3.974	41.015	4.407	0.802	1.339
HH	250	mod.2.same	-3.530	2.958	1.830	1.701	0.420	-0.635	-1.018	-1.683	13.516	1.972	0.649	0.755
HH	250	sev.1.same	0.415	18.201	14.749	14.381	11.066	8.296	5.638	2.075	51.451	5.696	0.321	0.658
HH	250	sev.1.switch	-1.490	13.888	10.954	10.735	7.808	4.424	2.263	1.202	54.378	5.083	0.579	1.735
HH	250	sev.1.under	0.901	20.530	15.191	15.627	8.086	3.621	2.976	2.146	70.524	8.371	0.413	0.594
HH	250	sev.2.under	-4.205	1.896	0.753	0.835	-0.569	-1.360	-1.735	-2.415	9.264	1.609	0.158	-0.312
HH	1000	mod.1.same	3.492	33.936	27.913	27.557	21.978	17.461	14.315	8.448	58.854	8.468	0.057	-0.270
HH	1000	mod.1.switch	5.306	24.397	21.042	20.785	17.424	14.737	13.151	10.644	42.694	5.052	0.261	-0.118
HH	1000	mod.2.same	-4.207	3.733	2.553	2.073	0.588	-0.633	-1.155	-2.255	17.531	3.001	1.369	2.601
HH	1000	sev.1.same	3.105	32.949	27.479	27.025	22.163	18.114	14.285	7.693	58.810	8.092	0.024	0.050
HH	1000	sev.1.switch	0.685	24.081	19.894	20.271	16.357	11.341	7.574	4.398	42.284	6.246	-0.344	0.133
HH	1000	sev.1.under	4.911	38.424	30.173	31.202	21.909	13.251	11.018	7.999	68.020	11.396	-0.105	-0.669
HH	1000	sev.2.under	-4.030	2.375	1.194	1.198	-0.077	-0.993	-1.508	-2.469	7.455	1.699	0.110	-0.279
HM	250	mod.1.same	-0.218	19.945	16.008	15.838	11.703	8.299	6.523	4.286	46.859	5.969	0.303	-0.026
HM	250	mod.1.switch	1.131	16.086	13.013	12.768	9.587	7.179	6.028	4.293	43.268	4.635	0.426	0.176
HM	250	mod.2.same	-4.519	3.014	1.639	1.610	-0.059	-1.019	-1.418	-2.080	13.035	2.097	0.451	0.161
HM	250	sev.1.same	1.582	20.127	16.151	15.963	11.950	8.623	6.479	3.875	44.850	5.933	0.233	-0.037
HM	250	sev.1.switch	-1.452	15.181	11.653	11.373	7.734	5.001	3.729	2.316	43.866	5.171	0.387	-0.043
HM	250	sev.1.under	1.371	19.346	15.201	14.568	10.037	6.859	5.519	3.933	64.068	6.889	0.817	1.428
HM	250	sev.2.under	-3.443	2.291	1.084	1.140	-0.324	-1.180	-1.541	-2.186	7.577	1.723	0.184	-0.480
HM	1000	mod.1.same	6.978	37.679	30.643	30.624	23.600	18.186	15.719	12.269	63.542	9.203	0.033	-0.714
HM	1000	mod.1.switch	5.266	29.203	24.117	23.720	18.767	15.494	13.780	10.821	47.566	6.807	0.172	-0.642
HM	1000	mod.2.same	-2.609	4.105	2.999	2.579	1.339	0.235	-0.357	-1.073	15.529	2.504	1.096	1.519
HM	1000	sev.1.same	7.215	37.813	30.992	30.703	24.212	19.653	16.994	12.863	60.853	8.780	0.050	-0.691
HM	1000	sev.1.switch	2.375	27.353	22.046	21.575	16.812	12.594	10.311	7.035	47.947	7.311	0.137	-0.512

Item Type	Sample Size	Miss. Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
HM	1000	sev.1.under	7.596	36.945	30.459	30.323	23.509	18.126	15.573	11.300	67.929	9.311	0.131	-0.375
HM	1000	sev.2.under	-3.132	3.535	2.106	1.928	0.555	-0.352	-0.805	-1.573	9.231	1.981	0.341	-0.446
HL	250	mod.1.same	0.668	20.019	16.691	16.653	13.294	9.638	7.543	5.047	48.125	5.361	0.210	0.406
HL	250	mod.1.switch	2.756	17.551	14.768	14.496	11.650	9.395	8.345	6.544	39.729	4.281	0.434	0.228
HL	250	mod.2.same	-3.655	3.171	1.892	1.871	0.422	-0.716	-1.150	-1.867	12.376	1.990	0.371	0.163
HL	250	sev.1.same	2.576	20.146	17.160	16.915	13.937	11.141	9.227	6.270	47.937	4.944	0.365	0.666
HL	250	sev.1.switch	1.003	16.965	13.734	13.675	10.330	7.315	5.867	4.092	37.117	4.822	0.218	-0.092
HL	250	sev.1.under	2.933	22.633	18.109	17.879	12.824	9.253	7.673	5.621	60.712	6.896	0.431	0.125
HL	250	sev.2.under	-3.819	2.608	1.475	1.423	0.187	-0.728	-1.176	-1.886	8.018	1.727	0.309	-0.132
HL	1000	mod.1.same	4.689	36.937	31.590	32.316	27.138	20.282	16.752	11.868	60.663	7.813	-0.422	0.028
HL	1000	mod.1.switch	10.622	32.551	28.138	28.053	23.597	20.131	18.257	15.395	51.105	6.123	0.100	-0.494
HL	1000	mod.2.same	-2.648	5.121	3.765	3.411	1.980	0.829	0.186	-0.807	16.938	2.564	0.795	0.715
HL	1000	sev.1.same	8.028	37.005	32.400	32.746	28.190	23.392	20.168	15.377	58.706	6.786	-0.249	0.069
HL	1000	sev.1.switch	4.789	31.573	26.302	26.803	21.348	16.277	13.604	9.549	48.851	7.215	-0.222	-0.416
HL	1000	sev.1.under	8.584	42.662	35.106	35.453	27.127	20.418	17.710	13.530	74.101	10.700	0.030	-0.565
HL	1000	sev.2.under	-2.257	4.065	2.630	2.252	0.974	0.052	-0.418	-1.011	11.934	2.195	0.616	-0.081
MH	250	mod.1.same	0.495	11.641	9.638	9.364	7.311	5.740	4.881	3.534	27.537	3.209	0.525	0.356
MH	250	mod.1.switch	-0.758	8.919	7.387	7.203	5.661	4.355	3.661	2.398	20.770	2.482	0.502	0.676
MH	250	mod.2.same	-2.880	2.444	1.505	1.403	0.404	-0.462	-0.783	-1.335	10.147	1.551	0.576	0.614
MH	250	sev.1.same	0.573	11.512	9.547	9.246	7.266	5.698	4.849	3.479	27.011	3.178	0.541	0.388
MH	250	sev.1.switch	-1.344	8.784	7.125	7.032	5.305	3.820	3.049	1.738	21.063	2.616	0.350	0.362
MH	250	sev.1.under	1.975	13.546	10.829	10.725	7.790	5.553	4.597	3.459	30.410	4.047	0.379	0.036
MH	250	sev.2.under	-2.470	1.453	0.687	0.694	-0.196	-0.776	-1.040	-1.468	6.114	1.120	0.242	-0.236
MH	1000	mod.1.same	5.883	21.470	18.219	17.800	14.691	12.419	11.226	9.347	38.585	4.666	0.332	-0.372
MH	1000	mod.1.switch	2.533	15.965	13.798	13.704	11.595	9.807	8.782	6.834	25.069	3.156	0.101	-0.161
MH	1000	mod.2.same	-2.256	3.812	2.654	2.350	1.083	-0.016	-0.515	-1.105	14.786	2.286	0.958	1.291
MH	1000	sev.1.same	5.142	21.172	18.083	17.568	14.709	12.543	11.359	9.397	37.919	4.558	0.394	-0.272
MH	1000	sev.1.switch	1.340	15.762	13.518	13.551	11.318	9.237	7.899	5.047	27.619	3.385	-0.089	0.126
MH	1000	sev.1.under	6.618	24.870	21.265	21.282	17.661	14.045	12.125	9.575	39.774	5.342	0.011	-0.293

Item Type	Sample Size	Miss. Type	Min	25%ile	Mean	Median	75%ile	90%ile	95%ile	99%ile	Max	SD	Skew	Kurt
MH	1000	sev.2.under	-3.154	1.926	1.003	1.052	0.054	-0.825	-1.257	-1.933	6.327	1.351	0.008	-0.259
MM	250	mod.1.same	0.954	12.430	10.318	10.031	7.940	6.282	5.427	3.987	28.140	3.291	0.479	0.226
MM	250	mod.1.switch	0.008	9.783	8.099	7.946	6.287	4.889	4.098	2.750	24.221	2.595	0.370	0.361
MM	250	mod.2.same	-2.477	2.488	1.574	1.499	0.533	-0.340	-0.690	-1.232	9.806	1.480	0.492	0.512
MM	250	sev.1.same	-17.904	12.477	10.403	10.100	8.032	6.462	5.603	4.179	27.486	3.271	0.374	0.952
MM	250	sev.1.switch	-15.813	9.316	7.636	7.477	5.772	4.376	3.559	2.161	20.670	2.663	0.312	0.606
MM	250	sev.1.under	-16.963	12.925	10.866	10.581	8.487	6.765	5.811	4.230	29.879	3.436	0.352	1.862
MM	250	sev.2.under	-2.629	1.694	0.929	0.918	0.092	-0.585	-0.875	-1.384	6.360	1.144	0.200	-0.129
MM	1000	mod.1.same	5.705	23.259	19.599	19.128	15.787	13.266	11.966	9.859	37.653	5.044	0.266	-0.553
MM	1000	mod.1.switch	3.912	17.790	15.265	15.129	12.640	10.634	9.508	7.465	28.256	3.637	0.138	-0.328
MM	1000	mod.2.same	-2.087	3.940	2.795	2.569	1.336	0.278	-0.283	-0.897	11.838	2.096	0.706	0.575
MM	1000	sev.1.same	5.678	23.443	19.760	19.235	16.017	13.720	12.499	10.422	37.016	4.867	0.272	-0.583
MM	1000	sev.1.switch	1.058	17.101	14.527	14.451	11.943	9.729	8.479	5.752	28.248	3.774	0.046	-0.149
MM	1000	sev.1.under	7.399	24.534	21.050	21.160	17.616	14.540	12.942	10.223	39.679	4.866	-0.028	-0.382
MM	1000	sev.2.under	-2.362	2.380	1.494	1.454	0.548	-0.289	-0.663	-1.299	7.517	1.367	0.276	0.054
ML	250	mod.1.same	0.649	12.194	10.317	10.147	8.233	6.615	5.667	4.060	31.537	3.011	0.389	0.407
ML	250	mod.1.switch	-0.141	10.664	9.078	8.981	7.363	5.992	5.216	3.850	22.728	2.478	0.286	0.238
ML	250	mod.2.same	-2.358	2.610	1.686	1.635	0.661	-0.229	-0.631	-1.191	9.568	1.480	0.418	0.433
ML	250	sev.1.same	-19.132	12.273	10.451	10.244	8.439	6.870	6.003	4.463	27.042	2.945	0.365	1.037
ML	250	sev.1.switch	-13.168	10.369	8.710	8.619	6.898	5.458	4.637	3.180	21.886	2.603	0.234	0.508
ML	250	sev.1.under	-25.022	14.453	12.127	12.002	9.625	7.511	6.408	4.766	29.106	3.640	0.142	1.540
ML	250	sev.2.under	-2.386	1.822	1.027	1.004	0.148	-0.543	-0.842	-1.341	6.899	1.205	0.286	-0.090
ML	1000	mod.1.same	4.438	22.679	19.611	19.569	16.583	13.801	12.202	9.704	38.238	4.435	0.018	-0.245
ML	1000	mod.1.switch	3.983	19.776	17.381	17.351	14.963	12.979	11.858	10.096	30.710	3.385	0.046	-0.310
ML	1000	mod.2.same	-2.183	4.423	3.200	3.020	1.761	0.687	0.079	-0.727	12.306	2.062	0.517	0.230
ML	1000	sev.1.same	4.863	22.681	19.830	19.757	16.905	14.462	13.082	10.760	36.646	4.164	0.091	-0.217
ML	1000	sev.1.switch	3.046	19.222	16.669	16.754	14.188	11.820	10.517	8.036	29.621	3.655	-0.111	-0.200
ML	1000	sev.1.under	6.919	27.051	23.112	23.420	19.108	15.244	13.617	11.109	43.129	5.605	-0.088	-0.489
ML	1000	sev.2.under	-2.383	2.854	1.889	1.798	0.824	-0.086	-0.555	-1.209	9.759	1.557	0.404	0.281

References

- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and test are measuring. *Applied Measurement in Education*, 7, 255-278.
- Adams, R. J., Wilson, M. R., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.
- Anderson, L. W. and Krathwohl, D. R. (eds.) (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Bandalos, D. L., and Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the Social Sciences* (pp. 93-114). New York: Routledge.
- Baumgartner, H. and Homburg, C. (2006). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13, 139-161.
- Beauducel, A., and Wittman, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1), 41-75.

- Bechger, T. M., Verstralen, H. H. F. M., and Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika*, 67, 123–136.
- Bentler, P.M., and Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Bentler, P. M., and Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, 47, 563-592.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (eds.) (1956). *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. New York: David McKay.
- Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Research and Methods*, 17, 303-316.
- Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.
- Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K. A. & Long, J. S. (Eds.) *Testing Structural Equation Models*. pp. 136–162. Beverly Hills, CA: Sage.
- Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician*, 36(3), 153-157.
- Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.

- Carmines, E. G., and McIver, J. P. (1981). Analyzing models with unobservable variables. In G. W. Bohrnstedt and E. F. Borgatta (Ed.), *Social Measurement: Current Issues*, (pp. 65-115), Beverly Hills: Sage Publications.
- Chou, C.- P. and Bentler, P. M. (2002). Model modification in structural equation modeling by imposing constraints. *Computational Statistics & Data Analysis*, 41, 271-287.
- Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., and Kirby, J. (2003). The finite sampling properties of the RMSEA: Point estimates and confidence intervals. *Sociological Methods and Research*, 32, 208-252.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.
- De Champlain, A. F. (1999). *An overview of nonlinear factor analysis and its relationship to item response theory*. Law School Admission Council Statistical Report 95-3. Newton, PA: Law School Admission Council.
- de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.
- Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, 41, 261–270.
- Embretson, S. E. and Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

- Fan, X., and Sivo, S. (2005). Evaluating the sensitivity and generalizability of SEM fit indices while controlling for severity of model misspecification. *Structural Equation Modeling*, 12 (3), 343-367.
- Fan, X., and Sivo, S. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42(3), 509-529.
- Fan, X., Thompson, B, and Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 56-83.
- Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement*, 34(1), 10-26.
- Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, 35(1), 67-82.
- Finney, S. J., and DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course*. Greenwich, CT: Information Age Publishing.
- Fraser, C., and McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.
- Gerbing, D. W., and Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen, & J.S. Long (eds.), *Testing structural equation models*. Newbury Park, CA: Sage.

- Gierl, M. J., and Mulvenon, S. (1995, April). *Evaluating the application of fit indices to structural equation models in educational research: A review of the literature from 1990 through 1994*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Guo, B., Perron, B., and Gillespie, D. F. (2008). A systematic review of structural equation modeling in social work research. *British Journal of Social Work*, 39(8), 1556-1574.
- Gushta, M. M., Yumoto, F., and Williams, A. (2009). *Separating Item Difficulty and Cognitive Complexity in Educational Achievement Testing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Harrell, L. M. (2009). *Accuracy of global fit indices as indicators of multidimensionality in multidimensional rasch analysis*. Unpublished Doctoral Dissertation.
- Hartig, J., and Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within- and between-item multidimensionality. *Journal of Psychology*, 216(2), 89-101.
- Henson, R.A., and Templin, J.L. (2006). *Implications of Q-matrix misspecification in cognitive diagnosis*. Unpublished manuscript.

- Heene, M., Hilbert, S., Draxler, C., Ziegler, M., and Buhner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16(3) 319-336.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Hu, L.-T., and Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 76-99). Thousand Oaks, CA: Sage.
- Hu, L., and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Hutchinson, S. R. (1998). The stability of post hoc model modifications in confirmatory factor analysis models. *Journal of Experimental Education*, 66, 361-380.
- Jackson, D. (2007). The Effect of the Number of Observations per Parameter in Misspecified Confirmatory Factor Analytic Models. *Structural Equation Modeling*, 14(1), 48-76.

- Jackson, D. L., Gillaspy, J. A., Jr and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods*, 14(1), 6-23.
- Janssen, R., and De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, 34, 245-26.
- Jöreskog, K.G., and Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the methods of maximum likelihood*. Chicago: National Educational Resources.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury, CA: Sage.
- Kamata, A. and Bauer, D. J. (2008). A note on the relationship between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15, 136-153.
- Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research*, 24, 285-305.
- Kaplan, D. (1990). Evaluation and modification of covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, 25, 137-155.
- Kaplan, D. (1991). On the modification and predictive validity of covariance structure models. *Quality & Quantity*, 25, 307-314.

- Kolen, M. J. , and Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice* 29(3), 8-14.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement*, 41(3), 205-237.
- Levy, R., and Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 64, 208-232.
- Li, Y., and Rupp, A. A. (2011). Performance of the $S\text{-}\chi^2$ statistic for full-information bifactor models. *Educational and Psychological Measurement*, 71(6), 986-1005.
- Linacre, J. M. (2011). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com
- Lord, F. (1952). *A Theory of Test Scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation.
- Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.
- Luecht, R. M. (1996). Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context, *Applied Psychological Measurement*. 20, 389-404.
- MacCallum, R. C. (1986). Specification searches in covariance structure analysis. *Psychological Bulletin*, 100, 107–120.

- Marsh, H. W., Balla, J. R. and McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 102, 391-410.
- Marsh, H. W., Hau, K.-T., and Wen, Z. (2004) In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling*, 11(3), 320-341.
- Marsh, H.W., and Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First-and higher order factor models and their invariance across groups. *Psychological Bulletin*, 97, 562-582.
- Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, 26, 51–71.
- McDonald, R. P. (1989) An index of goodness of fit based on noncentrality. *Journal of Classification*, 6, 97–103.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical data. *Journal of Educational Statistics*, 11, 3-31.
- Mulaik, S.A., James, L.R., Van Alstine, J., Bennet, N., Lind, S., and Stilwell, C.D. (1989), Evaluation of Goodness-of-Fit Indices for Structural Equation Models. *Psychological Bulletin*, 105(3), 430-45.

- Muthén, L. K. and Muthén, B. O. (1998-2001). *Mplus user's guide* (2nd ed). Los Angeles, CA: Authors.
- Muthén, L.K. and Muthén, B.O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén
- Muthén, L.K., and Muthén, B.O. (2007). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, B., du Toit, S. H. C., and Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript. University of California – Los Angeles.
- Orlando, M., and Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Orlando, M., and Thissen, D. (2003). Further investigation of the performance of $S-\chi^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289-298.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing* [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
- Raykov, T., and Marcoulides, G. A. (2001). Can There Be Infinitely Many Models Equivalent to a Given Structural Equation Model? *Structural Equation Modeling*, 8, 142-149.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.
- Reise, S. (1990). A comparison of item and person fit methods of assessing model fit in IRT. *Applied Psychological Measurement*, 42, 127-137.
- Revelle, W. (2011) *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, <http://personality-project.org/r/psych.manual.pdf>.
- Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Assessment: Theory, Methods, and Applications*. New York: Guilford
- Saris, W. E., Satorra, A. and van der Veld, W. (2009), Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling*, 16, 561-582
- Silvia, E. S. M., and MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, 23, 297-326.

- Sivo, S. A., Fan, X., Witta, E. L., and Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education*, 74, 267-288.
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54, 371-384.
- Steiger, J.H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT, Inc.
- Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling*, 7(2), 149-162.
- Steiger, H. H., and Lind, J. M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.
- Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen, & J. S. Long (eds.), *Testing structural equation models*. Newbury Park, CA: Sage.
- Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.

- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359), Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, Zetas for statistical pattern classification. *Applied Measurement in Education*, 9, 65-75.
- Tay, L. and Drasgow, F. (2012). Adjusting the Adjusted χ^2/df Ratio Statistic for Dichotomous Item Response Theory Analyses: Does the Model Fit? *Educational and Psychological Measurement*, 72(1), 1-18.
- Thissen, D. and Wainer, H. (Eds) (2001) *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thurber, R. S., Shinn, M. R., and Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, 31(4). 498-513.

- Wolfe, E. W., Hickey, D. T., and Kindfield, A. C. H. (2009). An application of the multidimensional random coefficients multinomial logit model to evaluating cognitive models of reasoning in genetics. *Journal of Applied Measurement*, 10, 196-207.
- Wu, M., and Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, 18(2), 93-113.
- Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). *ConQuest: Multi-aspect test software* [Computer software]. Melbourne: Australian Council for Educational Research.
- Ximénez, C. (2009). Recovery of weak factor loadings in confirmatory factor analysis under conditions of model misspecification. *Behavioral Research Methods*, 41, 1038-1052.
- Zenisky, A. L., Hambleton, R. K., and Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, 39, 291-309.
- Zhang, B., and Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, 68, 181-196.